


3 1761 10374380 3

12-001

GOVT



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743803>

12
-001



408

SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1993

•

VOLUME 19

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA



JUNE 1993 • VOLUME 19 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1993

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

July 1993

Price: Canada: \$35.00

United States: US\$42.00

Other Countries: US\$49.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	B.N. Chinnappa	C. Patrick
	G.J.C. Hole	D. Roy
	F. Mayda (Production Manager)	M.P. Singh
	R. Platek (Past Chairman)	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	L.-P. Rivest, <i>Laval University</i>
J.-C. Deville, <i>INSEE</i>	D.B. Rubin, <i>Harvard University</i>
D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
R.E. Fay, <i>U.S. Bureau of the Census</i>	C.-E. Särndal, <i>University of Montreal</i>
W.A. Fuller, <i>Iowa State University</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
J.F. Gentleman, <i>Statistics Canada</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	J. Sedransk, <i>State University of New York</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	C.M. Suchindran, <i>University of North Carolina</i>
D. Holt, <i>University of Southampton</i>	J. Waksberg, <i>Westat Inc.</i>
G. Kalton, <i>University of Michigan</i>	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

Assistant Editors

P. Lavallée, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$35 per year in Canada, US \$42 in the United States, and US \$49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada

Volume 19, Number 1, June 1993

CONTENTS

In This Issue	1
Record Linkage and Statistical Matching	
S. BARTLETT, D. KREWSKI, Y. WANG and J.M. ZIELINSKI Evaluation of Error Rates in Large Scale Computerized Record Linkage Studies	3
T.R. BELIN Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment	13
Y. THIBAudeau The Discrimination Power of Dependency Structures in Record Linkage	31
F. SCHEUREN and W.E. WINKLER Regression Analysis of Data Files that are Computer Matched	39
A.C. SINGH, H.J. MANTEL, M.D. KINACK and G. ROWE Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption	59
<hr/>	
M.A. HIDIROGLOU, J.D. DREW and G.B. GRAY A Framework for Measuring and Reducing Nonresponse in Surveys	81
R.P. TREDER and J. SEDRANSK Double Sampling for Stratification	95
R.J. CASADY and J.M. LEPKOWSKI Stratified Telephone Survey Designs	103
Z. OUYANG, H.T. SCHREUDER, T. MAX and M. WILLIAMS Poisson-Poisson and Binomial-Poisson Sampling in Forestry	115

In This Issue

This issue of *Survey Methodology* features a special section on **Record Linkage and Statistical Matching**. Special thanks are due to Fritz Scheuren for coordinating the editorial work of this special section. One or two papers which also deal with this topic, and which were too late to be included in this issue, may appear in a later issue.

In record linkage two datafiles are combined by linking records which refer to the same unit. The objective may be to create an enriched datafile containing variables from both of the source files, or it may be to identify records referring to common units. In situations where record linkage is not possible, statistical matching could be used to create an enriched datafile. A datafile created by statistical matching may contain synthetic records in the sense that variables obtained from the different data sources need not refer to the same unit; however, it is hoped that the matched file still accurately reflects statistical relationships among the variables.

Bartlett, Krewski, Wang and Zielinski discuss the advantages and disadvantages of record linkage in epidemiological studies. Record linkage methodology and methodological issues are reviewed and illustrated with examples of two large scale record linkage studies in epidemiology. Issues in the analysis of data from linked files are also reviewed.

Belin describes an experimental approach to the evaluation of alternative record linkage procedures. The approach is illustrated through a factorial experiment investigating the effect of such factors as the choice of matching variables, assignment of weights, and other factors. The experiment uses data from the 1988 U.S. census/post-enumeration survey dress rehearsal.

Thibaudeau considers an alternative to the commonly used conditional independence model for the probabilities of matches in different comparison fields. Data from the 1988 St. Louis census/post-enumeration survey dress rehearsal is used for illustration. It is found that the conditional independence model is reasonable for the true links; however, a hierarchical log-linear model with some interaction terms is used for the true nonlinks.

Scheuren and Winkler consider the analysis of data from linked files. In particular they consider the problem of regression of a dependent variable from one source file onto an independent variable from another source file. The approach taken is to estimate and correct for biases due to possibly incorrectly linked records. The approach works well if the probability of a match being a true link (and hence the biases in the regression estimation) can be well estimated. Some empirical results are presented.

The last paper in this special section, by Singh, Mantel, Kinack and Rowe, deals with statistical matching rather than record linkage. The authors develop methods of matching which use auxiliary data to avoid the conditional independence assumption. They also consider imposing categorical constraints so that the matched file agrees with appropriate marginal or conditional categorical distributions obtained from the source files or from auxiliary information. The main conclusion of an empirical evaluation is that the use of appropriate auxiliary information can considerably improve the quality of the matched file.

Hidiroglou, Drew and Gray present standards for the definitions of nonresponse to surveys that are being adopted at Statistics Canada. This will facilitate the analysis of global trends in nonresponse and better understanding of differences in nonresponse to different surveys. Factors affecting nonresponse and measures taken to reduce it are also discussed and nonresponse for two major Statistics Canada surveys is examined.

Treder and Sedransk compare simple random sampling and three allocation methods for double sampling. The three allocation methods are proportional, Rao's and optimal.

Casady and Lepkowski propose stratified telephone survey designs, based on commercial lists of telephone numbers, as alternatives to the widely used two stage random digit dialing procedure known as the Mitofsky-Waksberg technique. The efficiencies of various sampling schemes for this stratified design, simple random digit dialing and the Mitofsky-Waksberg procedure are compared.

Ouyang, Schreuder, Max and Williams consider the problem of estimation in Poisson-Poisson and binomial-Poisson sampling. A number of estimators of totals and standard errors are developed and empirically evaluated in the context of estimation of total volume of usable wood in a stand of trees.

Starting with this issue, *Survey Methodology* is changing to a larger page size. This larger size is less expensive to print and will allow *Survey Methodology* to reduce its continuing production deficit. We also took this opportunity to redesign the cover. I hope you like the result of our efforts.

Evaluation of Error Rates in Large Scale Computerized Record Linkage Studies

S. BARTLETT, D. KREWSKI, Y. WANG and J.M. ZIELINSKI¹

ABSTRACT

Matching records in different administrative data bases is a useful tool for conducting epidemiological studies to study relationships between environmental hazards and health status. With large data bases, sophisticated computerized record linkage algorithms can be used to evaluate the likelihood of a match between two records based on a comparison of one or more identifying variables for those records. Since matching errors are inevitable, consideration needs to be given to the effects of such errors on statistical inferences based on the linked files. This article provides an overview of record linkage methodology, and a discussion of the statistical issues associated with linkage errors.

KEY WORDS: Computerized record linkage; Canadian Farm Operators Study; National Dose Registry Mortality Study; Threshold selection.

1. INTRODUCTION

In recent years, there has been a trend in environmental epidemiology towards the use of existing administrative databases as sources of information for health studies (Howe and Spasoff 1986; Carpenter and Fair 1990). In general terms, this involves linking records of human exposure to environmental hazards with records on health status, often using computerized methods for matching individual records from different databases (Newcombe 1988).

Computerized record linkage (CRL) methods have recently been used to examine the mortality experience of over 326,000 farm operators in Canada in relation to farm practices (Jordan-Simpson *et al.* 1990). This study involved linking the Canadian Mortality Data Base (CMDB) with the 1971 Census of Population and the 1971 Census of Agriculture. Preliminary results based on 70,000 male farm operators in Saskatchewan have indicated that, although the cohort as a whole demonstrated no excess mortality for specific causes of death, there was some evidence of a dose-response relationship between mortality due to non-Hodkins lymphoma and acres sprayed with herbicides among farms less than 1,000 acres in size (Wigle *et al.* 1990).

Another ongoing large-scale study which involves record linkage is based on the National Dose Registry (NDR) of Canada. The NDR contains information on occupational exposures to ionizing radiation experienced by approximately 255,000 Canadians dating back to 1950. The NDR has recently been linked to the CMDB to investigate associations between exposure to ionizing radiation and cancer mortality (Ashmore *et al.* 1993).

A number of other health studies have been conducted by linking exposure data to the CMDB. Howe *et al.* (1987) determined significantly elevated lung cancer in uranium miners in the Northwest Territories. Significant associations were determined between lung cancer and diesel fumes and coal dust in a cohort study of male pensioners of the Canadian National Railway Company (Howe *et al.* 1983). Shannon *et al.* (1984) linked employment records of nickel workers in Ontario to the CMDB and found an excess in laryngeal and lung cancer mortality. Morrison *et al.* (1988) found significantly elevated risk of cancer of the lung, salivary gland, buccal cavity and pharynx among Newfoundland underground fluor spar miners. Mao *et al.* (1988) used CRL to link the CMDB to the Alberta Cancer Registry to determine survival rates after diagnosis for a wide range of cancers. The Canadian Labor Force Survey data base has been linked to the CMDB to examine the mortality experience of different occupations (Howe and Lindsay 1983). A comprehensive list of other health studies based on linking exposure data with the CMDB was compiled by Fair (1989).

Record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same individual. The procedures for CRL have become highly refined, using sophisticated algorithms to evaluate the likelihood of a correct match between two records (Hill 1988; Newcombe 1988). Statistics Canada has developed a CRL system called CANLINK which is capable of handling both one-file or internal linkages as well as linkages between two separate files (Howe and Lindsay 1981; Smith and Silins 1981).

¹ S. Bartlett, D. Krewski, Y. Wang and J.M. Zielinski, Environmental Health Directorate, Health Protection Branch, Health and Welfare Canada, Ottawa, Ontario, Canada K1A 0L2.

The confidentiality of records protected under the Statistics Act is strictly maintained if they are to be used in a study requiring record linkage. All studies requiring linkage with protected data bases must satisfy a rigorous review and approval process prior to implementation. All linked files with identifying information remain in the custody of Statistics Canada (Labossière 1986).

Record linkage studies have several advantages over traditional epidemiological studies. By using existing administrative databases, the need to collect new data for health studies is circumvented. By accessing existing data, large sample sizes can often be achieved with relatively little effort. Depending on the nature of the databases utilized, record linkage provides an inexpensive way of exploring many possible associations in epidemiological studies.

Record linkage also has a number of disadvantages. Matching errors may occur due to coding differences or nonuniqueness of the identifiers. There is generally little control over the information collected and there can be appreciable loss to follow-up. Record linkage studies also suffer from the same deficiencies as conventional epidemiological studies, including possible biases, confounding, and insensitivity to weak associations between the environment and health.

The purpose of this article is to explore the use of computerized record linkage in epidemiological studies based on administrative health and environmental records. Of particular interest is the impact of false links on statistical inferences about environmental health hazards. Algorithms for computerized record linkage are discussed in section 2. Applications of record linkage in studies of occupational exposure to ionizing radiation and agricultural chemicals are described in section 3. A discussion of statistical issues in

the analysis of data bases formed by record linkage is given in section 4. Our conclusions concerning the use of record linkage as a tool for use in environmental epidemiology are presented in section 5.

2. ISSUES IN RECORD LINKAGE

2.1 Problem Definition

Consider two computer files, **A** and **B**, consisting of health data and environmental exposure data, respectively, for two groups of individuals. Each file consists of a number of records or "observations", each containing a number of fields or "components". Typically, each observation corresponds to an individual member of the population. Fields are attributes such as name, address, age, and sex which characterize the observations. Record linkage is used to identify and link observations on each file that correspond to the same individual (Figure 1). In this example, record 1 of file A matches record 1 in file B, and record 2 in data base A matches record 3 in file B. Record 3 in file A does not match any records in file B, nor does record 2 in file B match any records in file A.

If the records contain unique identifiers which were accurately assigned, then the matching operation is trivial. The social insurance number is an example of an identifier that is unique to an individual. However, unique identifiers may not be available, in which case a "hard" linkage cannot be performed and thus some form of probabilistic linkage must be considered (see section 2.3). With this latter form of linkage, the likelihood of a correct match is computed, and a system of linkage weights is used to determine links and nonlinks.

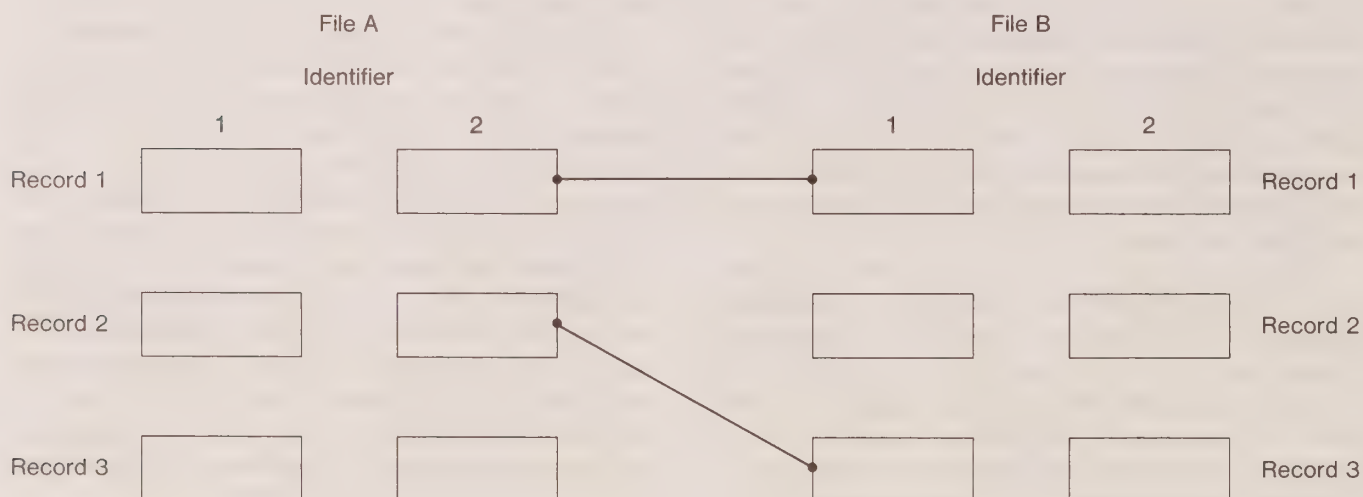


Figure 1. Schematic Diagram of Linking Two Files

2.2 Computerized Record Linkage System (CRL)

In probabilistic record linkage systems, the preliminary matching decision is based on a weight obtained from the comparisons of components of a pair of records (Newcombe 1988). The weight reflects the degree to which the pair is likely to be a true link: the higher the weight, the more likely the pair is a true link. The weight is commonly based on the odds in favour of a match when comparing two records,

$$\frac{P(M|AB\dots Z)}{P(\bar{M}|AB\dots Z)} = \frac{P(A|M)P(B|M)\dots P(Z|M)P(M)}{P(A|\bar{M})P(B|\bar{M})\dots P(Z|\bar{M})P(\bar{M})}.$$

Here, M is the event that two records match and $\{A, B, \dots, Z\}$ are outcomes of the comparisons of individual identifiers. The weight w is defined by the log-odds

$$w = \log_2 \left\{ \frac{P(M|AB\dots Z)}{P(\bar{M}|AB\dots Z)} \right\} \\ = W_a + W_b + \dots + W_z + W,$$

where

$$W_j = \log_2 \left\{ \frac{P(J|M)}{P(J|\bar{M})} \right\}$$

for all $J \in \{A, B, \dots, Z\}$, and

$$W = \log_2 \left\{ \frac{P(M)}{P(\bar{M})} \right\}.$$

It should be noted that in order to obtain an absolute odds, it is necessary to know the number of true matches and the number of non-matches. Otherwise, only the relative odds ratio can be determined. The weight determined by the CRL system used by Statistics Canada is the relative log-odds ratio.

Algorithms have been developed for assigning weights for the likelihood of a link between two records, based on the assumption that the likelihoods of the match for the individual identifiers are statistically independent (Howe and Lindsay 1981). Some identifiers, however, may be correlated leading to bias in the assignment of the overall weight.

Fellegi and Sunter (1969) proposed a mathematical model to provide a theoretical framework for record linkage. In the Fellegi-Sunter model, the weight takes into account the error probabilities for each field by using a likelihood ratio, with the weight w defined by

$$w = \sum_{i \in \{\text{fields}\}} w_i,$$

where

$$w_i = \begin{cases} \log_2\{m_i/u_i\} & \text{if field } i \text{ of a record pair agrees} \\ \log_2\{(1 - m_i)/(1 - u_i)\} & \text{if field } i \text{ of a record pair disagrees,} \end{cases}$$

with

$$m_i = \Pr\{\text{field } i \text{ agrees} \mid \text{record pair} \in M\} \quad (1)$$

and

$$u_i = \Pr\{\text{field } i \text{ agrees} \mid \text{record pair} \in U\}. \quad (2)$$

Here, M is a set of true matched record pairs and U is a set of un-matched pairs of records. The outcomes of each field comparison are also assumed to be statistically independent (Jaro 1989).

Newcombe (1988), Fellegi and Sunter (1969), Tepping (1968), Copas and Hilton (1990) developed various probabilistic and model-based approaches for assigning weights to components (fields) of records. A probabilistic system like the one used at Statistics Canada determines linkage weights by computing the logarithm of observed odds in favour of a match; other model-based systems use the EM algorithm (Dempster *et al.* 1977) to estimate linkage weights (Jaro 1989; Belin 1989; Winkler 1988).

2.3 Sources of Error

There are a number of sources of potential errors in record linkage that may lead to mismatching of records. Coding errors, such as the wrong birthdate, may occur when records are entered into data bases. There could be variations in the codes, such as different versions of the given name or surname.

In addition to coding errors and coding variations, missing data, especially for important identifiers, will significantly increase the error rate for record linkage (Fair and Lalonde 1988). Duplicate records, which occur when the same record in one file is matched with more than one record in the second file, could also lead to linkage errors (Jabine and Scheuren 1986). Because of this, CRL systems need to include rules that permit multiple matches.

One technique used for increasing the reliability of the surname identifier is to use a phonetic coding system. For example, two observations of an identifier, **ANDERSON** and **ANDERSEN**, will both be recoded as **ANDAR** by using the New York State Intelligence and Identification System (NYSIIS) (Newcombe 1988). Thus, the impact of variations in the name on the linkage would be minimized. However, in compressing the name, the power to discriminate between records may be diminished since two different names may have the same NYSIIS code. The likelihood of making incorrect links also increases (Newcombe 1988).

The given name may have variations with different versions entered on different data bases. Examples are William and Bill, Cynthia and Cindy, and David and Dave. Newcombe *et al.* (1992) discuss methods of using knowledge about variations in given names to increase the likelihood of a correct link.

Sometimes, the available identifiers may not adequately discriminate between individual records. The linkage algorithm may also under use the information contained in the the identifying fields used in the linkage process. Both situations can lead to matching errors.

For large files, it becomes impractical to compare all possible pairs of records. To reduce the number of comparisons, the records for the two files to be linked can be partitioned into mutually exclusive and exhaustive blocks and comparisons be made within blocks. Blocking is generally implemented by sorting the two files using one or more identifying variables. A disadvantage of doing this is that pairs of records, assigned to different blocks would not be compared, and hence would be classified as non-matching. The pairs to be compared would only be drawn from those records where the sorting variables agree. Thus, the number of false negative links would increase (Newcombe 1987; Jaro 1989). Good blocking variables are those based upon blocks that contain nearly the same number of records (Jaro 1989).

In most applications of the Fellegi-Sunter method, results of comparisons for different matching fields are assumed to be independent. Kelley (1986) performed simulation studies to investigate the robustness of the U.S. Census Bureau's linkage system against violations of the independence assumption. For certain populations and linkage variables, it was found that violation of the independence assumption can have an appreciable effect on the linkage error rates.

Newcombe *et al.* (1983) compared the accuracy of computerized matching with that of corresponding manual searches in an epidemiological follow-up study. They found that the computerized matching was more successful than the manual searches, and less likely to yield false links with records not related to the study population. In both approaches, accuracy was strongly dependent on the degree of personal identifying information available on the records being linked. Fair and Lalonde (1987) reached the same conclusion after examining the influence of the availability or non-availability of various identifiers on linkage error rates.

Schnatter *et al.* (1990) tested the adequacy of the CRL system used at Statistics Canada for correctly identifying deaths. Deaths known to have occurred in a cohort of 17,446 refinery and petroleum workers were compared to deaths determined through record linkage to the CMDB. Of the deaths occurring in Canada, 98% were detected by the CRL system.

2.4 Threshold Selection and Error Rate Estimation

After weights have been assigned to all potential matched pairs, a decision is made about the likelihood of the match being a true link. With the Fellegi-Sunter method, each weight is compared to upper and lower thresholds and a decision made as follows.

$$\text{Potential link} = \begin{cases} \text{a link} & \text{if } w \geq w_u \\ \text{a possible link} & \text{if } w_l < w < w_u \\ \text{a non-link} & \text{if } w \leq w_l. \end{cases}$$

Here, w_l and w_u are the lower and upper linkage thresholds, respectively, which ideally are selected to minimize the number of possible links, holding the two types of classification errors (true links classified as non-links and true nonlinks classified as links) at or below given levels.

Where feasible, any matches classified as possible links are resolved manually. Additional information may be used to aid in making decisions about possible links. In many applications, however, manual resolution is not practical, especially for linkages with a large number of possible links. In these situations, a single threshold, $w_t = w_l = w_u$, may be determined so that only two outcomes are possible. Those links with weights greater than w_t are declared links; those with weights less than w_t are declared non-links.

The choice of the threshold w_t is not straightforward. Existing methods are based on knowledge of the linkage error rates which are estimated either by manually resolving a sample of (if not all) possible links, or analytically. The former is a sample based approach since it involves the collection of data to estimate the linkage error rate.

The error rates for record linkage depend on how the thresholds are set. The larger the difference between the upper and lower thresholds, the more possible links there are. With a single threshold, the number of false negatives increases and the number of false positives decreases as the threshold increases.

A simple sample based approach for selecting the threshold entails a pilot study. First, a sample of the smaller of the two files to be linked is selected. Second, links are determined both manually and using a computerized probabilistic record linkage system. Third, assuming that the manually matched links are true links, the threshold is chosen as the weight at which the number of false positives plus the number of false negatives is minimized. Even so, linkage errors could still occur in the manually resolved links due to coding errors, insufficient discriminatory power in the identifiers used, or other linkage problems.

To estimate the error rates of a CRL system, a 2×2 contingency table can be constructed as follows.

CRL	Manual	
	Linked	Unlinked
Linked	n_{11}	n_{12}
Unlinked	n_{21}	n_{22}

The false positive (FP) and false negative (FN) rates are then estimated by

$$\text{FP} = \frac{n_{12}}{n_{11} + n_{12}}$$

and

$$\text{FN} = \frac{n_{21}}{n_{21} + n_{22}}.$$

Fellegi and Sunter (1969) point out that the error rates associated with given thresholds are functions of the agreement probabilities for true matches and true non-matches. Consequently, estimates of the agreement probabilities can be used to determine thresholds. This approach is also discussed by Jaro (1989).

For model-based record linkage systems, the principles for linking pairs of records and the strategy for setting thresholds are, with some modifications, similar to the sample based described above. The emphasis of this approach is to fit models for estimating conditional probabilities given by (1) and (2) and for estimating error rate using log odds of two estimated conditional probabilities. One such system uses the EM algorithm to estimate the conditional probabilities m_i and u_i given in (1) and (2) for the i th field of the record by assuming independence of the comparisons among fields,

$$\text{Pr}(\gamma^j | \mathbf{M}) = \prod_{i=1}^n m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j},$$

$$\text{Pr}(\gamma^j | \mathbf{U}) = \prod_{i=1}^n u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j},$$

($i = 1, \dots, n$ and $j = 1, \dots, N$), where n is the number of fields, N is the number of all comparison pairs, and

$$\gamma_i^j = \begin{cases} 1 & \text{if field } i \text{ agrees for record pair } j \\ 0 & \text{if field } i \text{ disagrees for record pair } j. \end{cases}$$

Iterating between the expectation step (E-step) and the maximization step (M-step) in the EM algorithm yields estimates of conditional probabilities \hat{m}_i and \hat{u}_i . The overall probability of correct matches may then be estimated based upon \hat{m}_i and \hat{u}_i (Jaro 1989).

Belin and Rubin (1991) provide a procedure which uses previous computer matching experience to fit a mixture

model for estimating the linkage error rate. A Box-Cox transformation (Box and Cox 1964) is applied to the weights for matches and for non-matches so that the transformed weights w_i^* form Gaussian distributions, φ_T and φ_F , with means μ_T and μ_F , and variances σ_T^2 and σ_F^2 , respectively. All transformed weights are then assumed to come from a mixture distribution

$$\lambda \varphi_T \left(\frac{w_i^* - \mu_T}{\sigma_T} \right) + (1 - \lambda) \varphi_F \left(\frac{w_i^* - \mu_F}{\sigma_F} \right).$$

After estimating the mixture coefficient λ using information obtained from previous matching experience, the above model can be fit using weights obtained from linkage procedure. The error rate for the record linkage algorithm given a particular threshold can then be estimated using the fitted model. The associated standard error of the estimated error rate is also estimated by using the SEM algorithm in which the covariance of the estimated parameters provided by the EM algorithm is estimated (Meng and Rubin 1991).

3. EXAMPLES OF LARGE RECORD STUDIES

3.1 Canadian Farm Operators Study

The Canadian Farm Operators Study was initiated to investigate possible relationships between causes of death in farm operators and various socio-demographic and farming variables. In particular, relationships between pesticide use and mortality are of interest. Mortality data was obtained from the CMDDB, while the socio-demographic and farming variables were obtained from the Census of Population and the Census of Agriculture. Since exposure to pesticides was not directly available in the census data bases, variables such as the number of acres sprayed for the control of insects or weeds and the cost of agricultural chemicals was used as surrogate information. The analysis file containing the pertinent information was constructed using probabilistic record linkage.

3.1.1 Cohort Definition

The cohort consists of all male farmers who met the definition for farm operator in the 1971 Census. A farm operator is defined as the person responsible for the daily decisions to be made about the operation of the farm. Farm operators are not necessarily owners, but could be tenants or hired managers. Only one operator was designated for each farm. A farm as determined in the 1971 census was an agricultural holding with one or more acres and with sales of agricultural products of \$50 or more. There were 326,000 male individuals who were classified as farm operators (Jordan-Simpson *et al.* 1990). The mortality experience of the cohort was followed up to 1987.

3.1.2 Record Linkage Methodology

The analysis file for the Canadian Farm Operator Study was formed as a result of three separate linkages, the last of which was the linkage of the farm operator cohort file to the CMDB. Before this linkage was done, the farm operator cohort file needed to be constructed.

Socio-demographic data was available from the 1971 Census of Population and information on farming practices was available from the 1971 Census of Agriculture. The Census of Population contains records for every individual in Canada and was collected in two versions, a short form and a long form. The long form asked for more information than the short form and was randomly administered to one third of the households. The Census of Agriculture was administered at all agricultural holdings.

Farm operators are not specifically identified by name in the Census of Agriculture file nor in the Census of Population file. The name and addresses of farm operators are contained in the Central Farm Registry which was created as a mailing list for agriculture questionnaires.

CMDB contains records for all registered deaths reported by the provinces and territories since 1950 and is stored in a standardized, computerized format under the custody of Statistics Canada (Smith and Newcombe 1982). The total number of death registrations on the CMDB from 1950 to 1987 is 5.9 million. The file contains identifying information, plus the date, place and underlying cause of death coded using the International Classification of Disease (ICD) code.

The Statistics Act protects the confidentiality of all records in the CMDB and the Census of Population and Census of Agriculture. As stated previously, all studies requiring linking with these data bases must satisfy a rigorous review and approval process prior to implementation and the resulting linked files with identifying information remain in the custody of Statistics Canada.

To form the analysis file, all the files described above were linked together in three phases.

- (a) **Follow-up.** The 1971 and the 1981 Central Farm Registers were linked using CRL to determine if farmers listed in 1971 were still alive in 1981. This information was added to the 1971 Central Farm Registry to increase the probability of linkage to the correct individual in the CMDB.
- (b) **Farm operator cohort data base.** The 1971 Central Farm Registry with the follow-up information was merged with the Census of Agriculture in order to add names to the cohort data base. This was necessary for linkage to the mortality data base. The resulting file was then linked to the Census of Population file using CRL to form the farm operator cohort data base.
- (c) **Analysis file.** The farm operator cohort data base was linked to the CMDB using CRL. The resulting file contained sociodemographic, exposure and death data and was then suitable for analysis.

3.1.3 Threshold Selection

Thresholds were required for each of the three linkages completed to form the analysis file and for linkages based on the short form and the long form. For the mortality linkage, Statistics Canada used a sample based procedure for setting thresholds. This procedure is illustrated for the final linkage of the farm operator cohort data base for those who filled out the short census form with the mortality data base.

A sample of approximately 10% of the short form records filled out by the cohort of farm operators was selected (Statistics Canada 1991a). The links were then determined in two ways, by using the Statistics Canada CRL and by manual resolution using information from death records. The results of the linkages were then compared assuming that the linkages determined by manual resolution were true linkages.

Numbers of false positives and false negatives at a series of link weight thresholds are shown in Figure 2 for the short form linkage. The threshold was selected to minimize the total number of false positives and false negatives, and occurs at a threshold value of 8. The false positive error rate is estimated to be $(36/453) \times 100 = 7.9\%$, while the false negative error rate is estimated to be $(38/20,847) \times 100 = 0.2\%$ leading to an overall error rate of 8.1% (Table 1).

Table 1
Comparisons of Linked and Unlinked Records
Using CLR and Manual Resolution Based on
a Sample of Census Records in the
Farm Operator's Study

Computerized Record Linkage (CLR)	Manual Resolution		Total
	Linked	Unlinked	
Short Form			
Linked	417	36	453
Unlinked	38	20,809	20,847
Total	455	20,845	21,300
Long Form			
Linked	286	13	299
Unlinked	15	18,498	18,513
Total	301	18,511	18,812

To illustrate the impact of additional identifying information, a similar table can be constructed for the long form records (Table 1). The false positive and false negative error rates are $(13/299) \times 100 = 4.3\%$ and $(15/18,511) \times 100 = 0.1\%$, respectively, for an overall rate of 4.4%. Thus, with more identifying information, the error rates can be reduced.

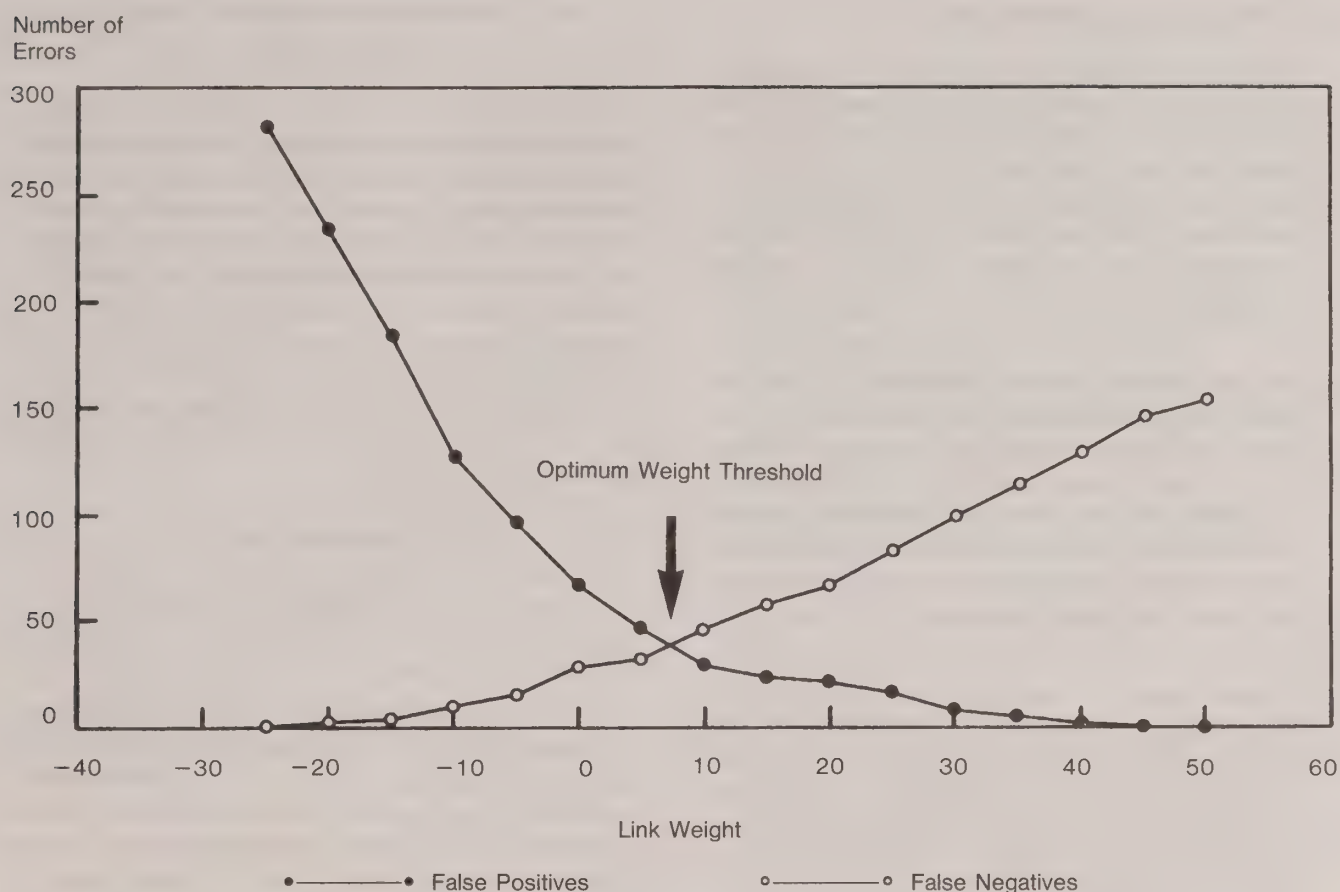


Figure 2. False Positive and False Negative Links Canadian Farmer Operators Mortality Linkage: Short Form

3.2 National Dose Registry Mortality Study

The National Dose Registry of Canada contains records of occupational radiation exposures for approximately 255,000 Canadians dating back to 1951. The NDR has recently been linked to the CMDDB. The purpose of the National Dose Registry mortality study is to determine associations between excess mortality due to cancer and other causes and occupational exposure to low levels of ionizing radiation (Ashmore *et al.* 1993).

3.2.1 Cohort Definition

The cohort consists of all workers monitored for ionizing radiation, including tritium and radon daughters, whose records were contained in the National Dose Registry as of December 31, 1983. It contains radiation exposure records of virtually all monitored radiation workers in Canada, with some records providing 37 years of exposure data. In addition, the Registry includes 80 different job categories ranging from nuclear power generating station workers to hospital radiologists to dentists. A total 248,940 people were included in the study cohort.

Depending on the type of radiation and the levels of exposure anticipated within specific job categories, radiation exposure records have been collected annually, quarterly, monthly, or biweekly. Each year, a summary measure of the annual dose experienced by each individual is recorded in the Lifetime Dose History System (LDHS). The annual exposure records maintained in the LDHS will be used as the basis for examining potential relationships between occupational radiation exposure and health status.

The individual data in the LDHS also permit the calculation of a cumulative lifetime dose for each individual. Although individuals will not experience the same level of exposure each year, an average annual dose for an individual can be obtained by dividing the cumulative lifetime dose by the number of years which have elapsed since the time of first exposure. Statistical analysis can be based on the cumulative lifetime dose, the average annual dose, or the annual doses as recorded on the LDHS. Up to 1986, personal identifying information such as surname, given name, sex, year of birth, and assigned identification numbers used to identify the individuals' dose records were stored separately in the Master Identification File (MIF) (Ashmore and Grogan 1985).

3.2.2 Record Linkage Methodology

Identifying variables changed form a number of times during the history of the NDR making tracing an individual's dose history difficult, at times. Because of these and related problems, the Social Insurance Number has been used as the key to the individuals' records from 1977 onward.

There were several linkages required to bring together the appropriate personal identifiers, dose histories, and death information.

- (a) **Dose history linkage.** Since 1984 Statistics Canada has been conducting dynamic merges to their LDHS database in order to regroup dose records; reducing the number of fragmented records and consolidating the records into comprehensive dose histories for each study member. The file resulting from the internal linkages indicated which records on the NDR appear to belong to the same individual.
- (b) **CMDB linkage.** The internally linked MIF cohort was linked to the mortality records (two-file linkage). By linking the two, it is possible to measure the cohort members' subsequent risk of death. In this study, the CMDB was used to obtain the underlying cause, year of death, the place of death, place of birth, and birth year information.
- (c) **Analysis file.** A match of data from MIF, the CMDB and the LDHS was performed to create a comprehensive record for each member of the study cohort. Where the information is available, each record includes birth month and year, sex, the death data listed above, the death linkage weight, and a dose history. Any unmatched records from the MIF or dose history file have undergone special scrutiny.

3.2.3 Threshold Selection

Threshold selection for the link of the cohort file to the CMDB was done in a manner similar to that used in the Canadian Farm Operator Study. First, the weights of potential links were determined. All potential links that had weights less than -30 were considered to be nonlinks. There were 4,429 female and 8,686 cohort members with linkage weights above this value. A sample of these remaining individuals was selected and manually resolved by reviewing death certificates to determine if the links were true links or nonlinks. The threshold was selected at the link weight for which the number of false positive links was equal to the number of false negative links for females and males separately. For females the selected threshold was 53 and for males, 27 (figure not shown).

4. ISSUES IN ANALYSIS OF LINKED DATA SETS

Relatively little work has been done to determine the impact of record linkage on the results of regression analysis. Neter *et al.* (1965) recognized that errors introduced during the matching process could adversely affect analysis based on the resultant linked files. Suppose that true values of a random variable of interest are recorded on a data file comprised of N records. Let Y_i denote the true value for record $i = 1, \dots, N$. This file is linked to a second file containing identifying information, following which a value of Z_i is assigned to record $i = 1, \dots, N$. Assuming that all matching errors are equally probable, we have,

$$Z_i = \begin{cases} Y_i & \text{with probability } p \\ Y_j & \text{with probability } q(j \neq i), \end{cases}$$

where $p + (N - 1)q = 1$.

Neter *et al.* (1965) used this model to study the impact of matching errors on the sample mean and variance of the variable Z . The effect of matching errors on the correlation between Z and a second random variable X contained on the same file as well on parameter estimates of the regression between Z_i and X_i , was also investigated. It was shown that (1) the estimate of the mean of Z is unbiased for the mean of the Y ; (2) if " X " is positively correlated with Y , the residual variance from a regression of Z on X will be larger than the variance from a regression of Y on " X "; and (3) the slope of the regression line will be underestimated when Z is used rather than Y .

Belin and Rubin (1991) and Winkler and Thibaudeau (1991) discuss theoretical framework, computational algorithms, and software for estimating matching probabilities. These advances motivated Scheuren and Winkler (1991) to update the work of Neter *et al.* (1965). They used the model

$$Z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij}(j \neq i), \end{cases}$$

where $p_i + \sum_{j \neq i} q_{ij} = 1$, to study the impact of matching errors on the estimates of the coefficients β in the linear regression model

$$Y = X\beta + \epsilon.$$

The effect of matching errors on the above regression model may be expressed as

$$E(Z_i) = Y_i + B_i,$$

where the bias term is given by

$$B_i = (p_i - 1)Y_i + \sum_{j \neq i} q_{ij} Y_j.$$

Instead of using the pair of independent and dependent variables (X_i, Y_i) , the pair of independent and linked dependent variables (X_i, Z_i) is used to fit the model. Noticing that the linked dependent variable may be written as $Z = Y + B$, the coefficients are estimated as

$$\hat{C} = (X^T X)^{-1} X^T Z = \hat{\beta} + (X^T X)^{-1} X^T B,$$

where $\hat{\beta} = (X^T X)^{-1} X^T Y$, so that the bias adjustment is $(X^T X)^{-1} X^T B$.

Scheuren and Winkler (1991) used these models conduct simulation studies based on real data. Their approach was to take a file of linked and nonlinked cases and re-link them using different matching variables. This simulation demonstrates that the ability to accurately estimate matching probabilities critically effects the accuracy of the coefficient estimates. If the matching probabilities can be accurately estimated, the adjustment procedure works reasonably well.

5. DISCUSSION

Record linkage provides an attractive methodology for exploring relationships between exposures and health outcomes by making use of existing data bases. However, linkage errors are possible, resulting from coding errors, variations in identifiers, missing data, and insufficient discrimination power in the identifiers.

Error rates depend on the amount of identifying information, as seen in the farm operators study. Here, the error rate decreased for the linkage of the long census form where more identifying information was available than in the short census form. Thus, it is important that good identifying information be available for record linkage.

Relatively little attention has been paid to the impact of linkage errors on statistical inferences based on record linkage studies. Such errors can lead to biases in estimates of measures of association between health and environmental variables, such as regression coefficients. Work is in progress to investigate the impact of these errors on the results of the epidemiological studies presented in this paper.

ACKNOWLEDGMENTS

We would like to thank Martha Fair, Statistics Canada, and Dr. Howard Morrison, Health and Welfare Canada for their helpful comments on this article. We also thank two anonymous reviewers for many constructive suggestions.

REFERENCES

- ASHMORE, J.P., and GROGAN, D. (1985). The National Dose Registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.
- ASHMORE, J.P., KREWSKI, D., and ZIELINSKI, J.M. (1993). National Dose Registry Study. *European Journal of Cancer*, submitted.
- BELIN, T.R. (1989). Results from evaluation of computer matching. Memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.
- BOX, G., and COX, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26, 211-246.
- CARPENTER, M., and FAIR, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference - 1989: Proceedings of Record Linkage Sessions and Workshop*. Ottawa, Ontario: Ottawa Select Printing.
- COPAS, J.B., and HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society*, Series A, 153, 287-320.
- DEMPSTER, A.D., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via EM algorithm, (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.
- FAIR, M.E. (1989). Studies and references relating to uses of the Canadian Mortality Data Base. Report from the Occupational and Environmental Health Research Unit, Health Division, Statistics Canada, Ottawa.
- FAIR, M.E., and LALONDE, P. (1988). Missing identifiers and the accuracy of individual follow-up. *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada, Ottawa, 95-107.
- FELLEGI, I.P., and SÜNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- HILL, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy. Release 2.7. Report from Research and General Systems, Informatics Services and Development Division, Statistics Canada, Ottawa.
- HOWE, G.R., and LINDSAY, J. (1981). A Generalized Iterative Record Linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.
- HOWE, G.R., and LINDSAY, J. (1983). A follow-up study of a ten-percent sample of the Canadian Labor Force. I. Cancer mortality in males, 1965-73. *Journal of the National Cancer Institute*, 70, 37-44.
- HOWE, G.R., FRASER, D., LINDSAY, J., PRESNAL, B., and YU, S.Z. (1983). Cancer mortality (1965-77) in relation to diesel fume and coal exposure in a cohort of retired railway workers. *Journal of the National Cancer Institute*, 70, 1015-1019.

- HOWE, G.R., NAIR, R.C., NEWCOMBE, H.B., MILLER, A.B., BURCH, J.D., and ABBATT, J.D. (1987). Lung cancer mortality (1950-80) in relation to radon daughter exposure in a cohort of workers at the Eldorado Port radium uranium mine: Possible modification of risk by exposure rate. *Journal of the National Cancer Institute*, 79, 1255-1260.
- HOWE, G.R., and SPASOFF, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. Toronto: University of Toronto Press.
- JARO, M.A., (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistical*, 2, 255-277.
- JORDAN-SIMPSON, D.A., FAIR, M.E., and POLIQUIN, C. (1990). Canadian Farm Operator Study: Methodology. *Health Reports*. Catalogue 82-003, Statistics Canada, 2, 141-155.
- KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. Paper Presented at the August 1986 meeting of the American Statistical Association.
- LABOSSIERE, G. (1986). Confidentiality and access to data: the practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press.
- MAO, Y., SEMENCIW, R., MORRISON, H., KOCH, M., HILL, G., FAIR, M., and WIGLE, D. (1988). Survival rates among patients with cancer in Alberta in 1974-78. *Canadian Medical Association Journal*, 138, 1107-1113.
- MENG, L., and RUBIN, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899-911.
- MORRISON, H.I., SEMENCIW, R.W., MAO, Y., and WIGLE, D.T. (1988). Cancer mortality among a group of fluorspar miners exposed to radon progeny. *American Journal of Epidemiology*, 128, 1266-1275.
- NETER, J., MAYNES, E.S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., and ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medicine*, 13, 157-169.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: methods for health and statistical studies, administration and business*. Oxford: Oxford Medical Publications.
- NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1204.
- SCHEUREN, F., and WINKLER, W.E. (1991). An error model for regression analysis of data files that are computer matched. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.
- SCHNATTER, A.R., ACQUIVELLA, J.F., THOMPSON, F.S., DONALESKI, D., and THERIAULT, G. (1990). An analysis of death ascertainment and follow-up through Statistics Canada's Mortality Data Base system. *Canadian Journal of Public Health*, 81, 60-65.
- SHANNON, H.S., JULIAN, J.A., and ROBERTS, R.S. (1984). A mortality study of 11,500 nickel workers. *Journal of the National Cancer Institute*, 73, 1251-1258.
- SMITH, M.E., and NEWCOMBE, H.B. (1982). Use of the Canadian Mortality Data Base for epidemiological follow-up. *Canadian Journal of Public Health*, 73, 39-46.
- SMITH, M.E., and SILINS, J. (1981). Generalized Iterative Record Linkage System. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.
- STATISTICS CANADA, (1991a). Canadian Farm Operators' mortality study general work plan. Internal report of the Occupational and Environmental Health Research Section, Statistics Canada, Ottawa.
- STATISTICS CANADA, (1991b). Unpublished data.
- TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- WIGLE, D.T., SEMENCIW, R.M., WILKINS, K., RIEDEL, D., RITTER, L., MORRISON, H.I., and MAO, Y. (1990). Mortality study of Canadian male farm operators: Non-Hodgkin's lymphoma mortality and agriculture practices in Saskatchewan. *Journal of the National Cancer Institute*, 82, 575-581.
- WINKLER, W.E. (1988). Using the E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.
- WINKLER, W.E., and THIBAUDEAU, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division, Technical Report.

Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment

THOMAS R. BELIN¹

ABSTRACT

Record linkage refers to the use of an algorithmic technique for identifying pairs of records in separate data files that correspond to the same individual. This paper discusses a framework for evaluating sources of variation in record linkage based on viewing the procedure as a “black box” that takes input data and produces output (a set of declared matched pairs) that has certain properties. We illustrate the idea with a factorial experiment using census/post-enumeration survey data to assess the influence of a variety of factors thought to affect the accuracy of the procedure. The evaluation of record linkage becomes a standard statistical problem using this experimental framework. The investigation provides answers to several research questions, and it is argued that taking an experimental approach similar to that offered here is essential if progress is to be made in understanding the factors that contribute to the error properties of record-linkage procedures.

KEY WORDS: Cutoff weight; False-match rate; Fellegi-Sunter algorithm; Matching variables; Post-enumeration survey; String comparison; Weighting scheme.

1. EVALUATING RECORD-LINKAGE PROCEDURES

Record linkage refers to the use of an algorithmic technique to identify pairs of records, one from each of two data files, that correspond to the same individual. The goal is to identify, using a computerized approach, the records from the respective data files that should be declared “matched” as well as the records that should be declared “not matched” without an excessive rate of error, thereby avoiding the cost of manual processing.

Specifying a record-linkage procedure requires both a method for measuring closeness of agreement between records and a rule for deciding when to classify records as matches or non-matches. Much attention has been paid in the record-linkage literature to the problem of assigning so-called “weights” to individual fields of information in a multivariate record to obtain a “composite weight” that summarizes the closeness of agreement between two individuals (*e.g.*, Newcombe *et al.* 1959; Fellegi and Sunter 1969; Newcombe 1988; Copas and Hilton 1990). Less attention has been paid to other aspects of record-linkage procedures, such as the handling of close but inexact agreement between fields of information, and to the effects of using various approaches (treatments) in combination with one another.

In some settings, a personal identifier, such as a social security number, can serve as a basis for linkage. However, such an identifier is not always available, and even when one is present, it still may be necessary to rely on other identifying information for a substantial subset of cases (*e.g.*, Rogot, Sorlie and Johnson 1986).

This paper describes a large factorial experiment contrasting various procedures for matching census and post-enumeration survey (PES) records. Social security number is not collected in the census, so we are in a setting where closeness of agreement is based on several variables. Interest focuses on two questions:

- (1) What are the most important factors affecting the accuracy of record linkage?
- (2) What combination of factors works best in practice?

Beyond addressing these questions in the census/PES setting, perhaps the most important contribution of this investigation is the idea that record-linkage procedures should be studied by conducting careful experiments. With many factors at the discretion of the operator of the program, there is little hope of understanding the full complexities of a matching algorithm by varying factors one at a time (or worse, not even conducting any systematic evaluation at all). The idea of conducting an experiment would seem quite natural to an agricultural scientist or an industrial quality-control engineer, although it seems that such an approach has not been taken in the context of record linkage aside from this investigation and earlier work by the author (Belin 1989a, 1989b).

2. APPLIED CONTEXT FOR RECORD LINKAGE

2.1 Applications of Record Linkage

Record-linkage methods have been used in a variety of settings. Applications can be characterized as falling into two broad groups: problems where it is desired to draw

¹ Thomas R. Belin, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA, 90024-1766, U.S.A.

inferences about relationships between variables collected in separate large data files, and problems where interest focuses directly on the number of individuals represented in one or both data files (or a function of those quantities).

Examples of the first type of application are numerous. Studies have been conducted linking data from health and nutrition surveys to registries of mortality data to study relationships between dietary risk factors and death from various causes (Johansen 1986), linking labor force survey data to mortality data to assess health effects of uranium mining (Newcombe, Smith, Howe, Mingay, Strugnell and Abbatt 1983; Abbatt 1986), linking information on educational background to records of earnings of individuals some years later to assess the benefit of a college education (Fagerlind 1975), comparing reported income on welfare records to reported income on tax records (Kershaw and Fair 1979), and linking records of individuals exposed to radiation during atomic-bomb tests and records of a cohort of control individuals to national death records to assess differences in mortality patterns between exposed and control individuals (Dulberg, Spasoff and Raman 1986). Using record-linkage methodologies in such studies is attractive primarily for reasons of cost and timeliness, since for any of the research endeavors just described, it would take much longer and would have been much more expensive to conduct studies with one or more stages of followup than it was to make use of existing data.

The primary motivating example in this article is representative of the other type of application, where the goal is to determine the number of overlapping cases in two data files. In this example, a record-linkage procedure is used as the first step of an extensive matching operation in which records from a census are compared to records from a large-scale post-enumeration survey (PES) conducted after the census to evaluate census coverage. Other examples where the goal is to determine the number of overlapping cases between data files are the investigation by Nicholl (1986) of classification errors regarding the types of injuries sustained by road accident victims (based on linking hospital records to police reports of accidents), the investigation by Johnson (1991) into caseloads for U.S. Attorneys in different districts around the country (based on linking a list of cases assembled by the Department of Justice to a list of cases assembled by federal district courts), and a variety of investigations into the accuracy and coverage of mortality data files (Wentworth *et al.* 1983; Curb *et al.* 1985; Boyle and Decouflé 1990; Williams *et al.* 1992).

Census undercount estimation has been a prominent and at times controversial topic in statistical research, especially during the past decade. Much of the controversy revolves around a proposed adjustment of the census based on undercount estimates from a PES. For general background on issues involved in census undercount

estimation, see Ericksen and Kadane (1985), Citro and Cohen (1985), Freedman and Navidi (1986), Wolter (1986), Schirm and Preston (1987), Ericksen, Kadane, and Tukey (1989), Cohen (1990), and the special sections on census coverage error in the June and December, 1988, issues of this journal. A record-linkage procedure is the first step of matching census records to PES records; it is followed by matching of records by clerks, subsequent followup interviewing of households when there appear to be discrepancies between the census and PES findings, and an additional round of clerical matching after followup interviewing. Based on assessments from the matching operation and certain assumptions about the probability that individuals would be included only in the census, only in the PES, in both the census and PES, or in neither the census nor PES, it is possible to estimate undercount (or overcount) rates in the census.

2.2 Background on Record-Linkage Theory

The development probabilistic reasoning in record-linkage theory can be traced to Newcombe, Kennedy, Axford, and James (1959), who develop a weighting scheme in an effort to reflect the odds that a pair of records is correctly matched. Fellegi and Sunter (1969) enhance the theoretical underpinnings of commonly-used weighting rules, noting that the procedure proposed by Newcombe *et al.*, corresponds to calculating a likelihood ratio under a simple model for the record-linkage problem that supposes independence of agreement among all fields of information within records. They show that a weighting scheme similar to that of Newcombe *et al.*, combined with cutoff weights that depend on a specified false-match rate and a specified false non-match rate, define a linkage procedure that is optimal in the sense of minimizing the proportion of records that will be assigned neither as definitely matched nor as definitely not matched, assuming the underlying model is valid.

Much of the ensuing development of record-linkage technology has taken place in the context of applications, as investigators put the theoretical ideas outlined in the earlier literature to practical use. Prominent applications include the Oxford Record Linkage Study (Acheson 1967; Goldacre 1986); the three-way match among records from the Current Population Survey, the Social Security Administration, and the Internal Revenue Service (Kilss and Scheuren 1978); and the National Longitudinal Mortality Study (Rogot, Sorlie, Johnson, Glover and Treasure 1988). The proceedings volumes from conferences on record linkage (Kilss and Alvey 1985; Howe and Spasoff 1986; Carpenter and Fair 1990), compilations of papers from annual conferences (Kilss and Alvey 1984a; Kilss and Alvey 1984b; Kilss and Alvey 1984c; Kilss and Alvey 1987; Kilss and Jamerson 1990), and proceedings volumes from conferences more broadly focused on uses of administrative

data (Coombs and Singh 1988) document numerous other applications that make use of record-linkage methodology.

Software development has enhanced the ability to pursue research into record linkage. Software incorporating refinements of weighting methods and blocking strategies has been developed for use in a variety of applications at Statistics Canada and the U.S. Bureau of the Census. Background on the Statistics Canada "Generalized Iterative Record Linkage System" (GIRLS) is discussed in Howe and Lindsay (1981); documentation is contained in Hill (1981) and Hill and Pring-Mill (1986). Background on the matching system developed by the Record Linkage Staff at the U.S. Bureau of the Census can be found in Jaro (1989), Winkler (1989), and Winkler and Thibaudeau (1992), with documentation found in Laplant (1988), Laplant (1989), and Winkler (1991).

New models that reflect subtleties within data files that could be used in developing a probabilistic weighting scheme are offered by Copas and Hilton (1990). Other extensions to record-linkage methodology designed to take advantage of information in person names are described in Newcombe, Fair and Lalonde (1992). A review paper by Jabine and Scheuren (1986), a textbook by Newcombe (1988), and a compilation by Baldwin, Acheson and Graham (1987) serve as broad references on record-linkage methodology.

2.3 Flow of a Standard Record-Linkage Procedure

Typical steps in a record linkage procedure can be described as follows: (1) data collection, (2) preprocessing of data, (3) determination of rules for assessing closeness of agreement between candidate matched pairs, (4) assignment of candidate matched pairs, and (5) declaration of matched pairs. We use the term "candidate matched pairs" to describe pairs of records that are brought together as being the best potential match for each other from the respective data files (*cf.* "hits" in Rogot, Sorlie, and Johnson (1986); "pairs" in Winkler (1989); "assigned pairs" in Jaro (1989)). Candidate matched pairs might be declared matched after the application of a decision rule in step (5), but they will not necessarily be declared matched by the decision rule.

As indicated earlier, closeness of agreement between candidate matched pairs is assessed in many record-linkage procedures by a univariate summary statistic, often referred to as a "composite weight". In such procedures, step (3) above would refer to the determination of weighting rules, and step (5) above would involve the setting of a cutoff weight above which record pairs will be declared matched.

Record linkage may be viewed as a decision problem with two or more actions to be taken by the computer. Typically, three actions are considered (*e.g.*, declare records matched, declare records as not matched, or send

record to be reviewed more closely by a human observer, as in Fellegi and Sunter 1969), although sometimes only two actions (declare matched, declare not matched) are contemplated, and as many as five actions have been considered in some instances (Tepping 1968).

Postulating that distance between multivariate records can be summarized by a univariate composite weight narrows the scope of possible procedures that could be used to perform record linkage. The author is aware of very little research exploring alternatives to such univariate-composite-weight approaches, other than merely specifying a deterministic set of rules for when to declare records matched; one exception is Smith and Newcombe (1975). Such alternatives are beyond the scope of this paper.

2.4 Detailed Description of the Procedure Used to Match Census/PES Records

A variety of separate techniques may be involved in each of the five steps outlined above. Figure 1 provides a flowchart illustration of the main steps used in the linkage of census/PES records.

The frame of the census is a compilation of housing-unit address listings. Addresses are assembled by a variety of techniques, generally depending on whether the area is urban or rural. In urban and suburban areas, census forms are mailed to households with the hope that residents will respond by mailing back a completed form; in other areas census enumerators visit households. When there is no response from a household that was sent a census form by mail, an enumerator will visit the household in person. Data are entered into Census Bureau computer files by a combination of computerized scanning techniques and clerical keying operations. An overview of census methodology can be found in Citro and Cohen (1985); detailed descriptions of various census operations can be found in the Census Bureau's 1990 Decennial Census Information Memorandum Series (Bureau of the Census 1988-1991).

Data collection in the type of post-enumeration survey conducted in 1990 (and in test censuses leading up to the 1990 PES) begins with a process of listing addresses that is conducted by enumerators canvassing neighborhoods. Information is obtained entirely through interviewing operations as opposed to the mailout-mailback approach. Data are entered into computer files entirely by clerical keypunching. Hogan (1992) provides an overview of the PES; details of PES operations can be found in the Census Bureau's STSD Decennial Census Memorandum Series (Bureau of the Census 1987-1991).

Preprocessing of data is rarely discussed in the literature on record linkage, even though this stage provides opportunities both for squeezing available information from the data at hand and for unwisely discarding information available from the data. Winkler (1985a, 1985b) presents

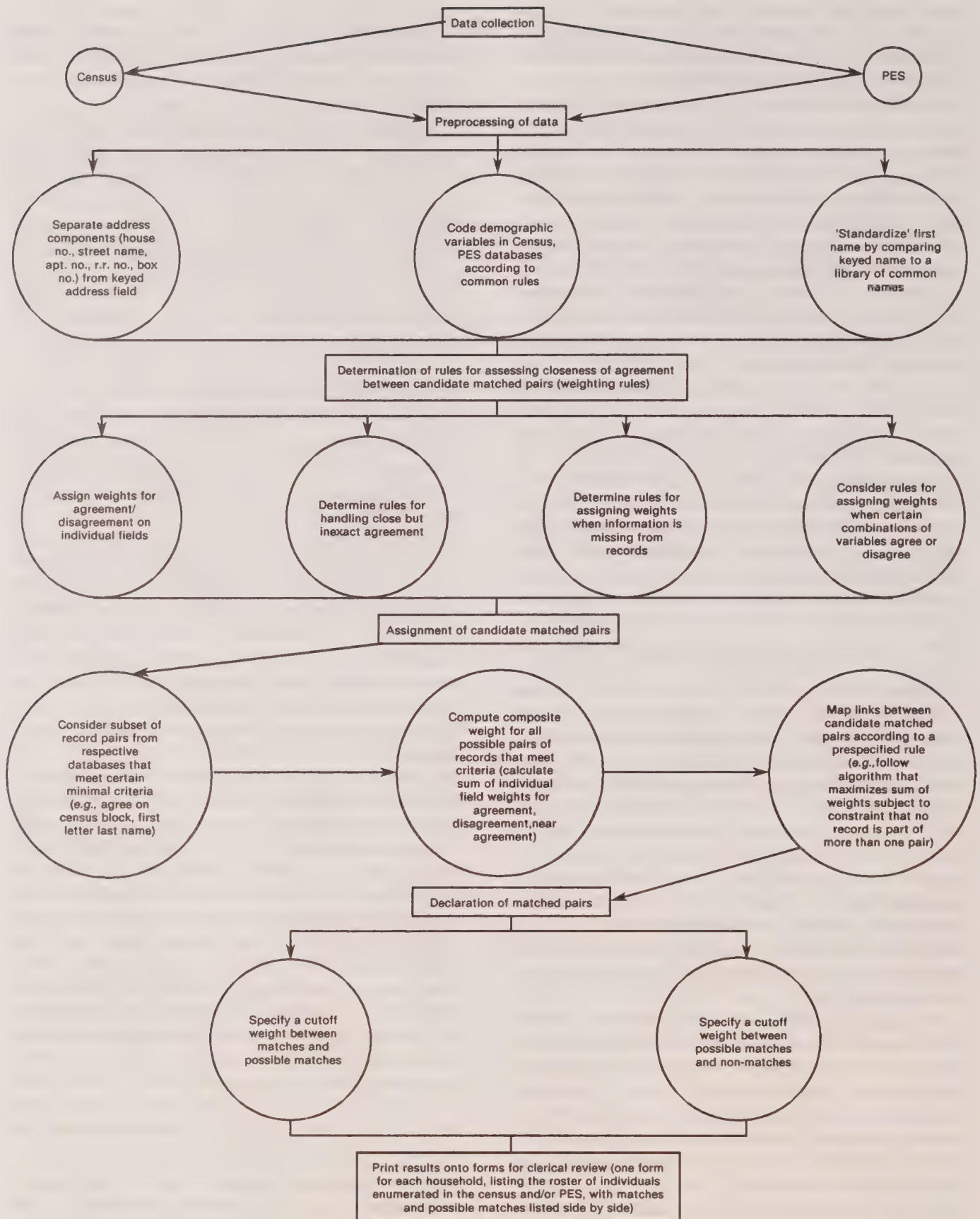


Figure 1. Flowchart of Census/PES Record Linkage Procedures

some specific strategies that are shown to make it easier to distinguish true matches from false matches, and Jabine and Scheuren (1986) and Newcombe (1988) offer some broad guidelines in this area. In the census/PES matching operation, preprocessing of data includes coding demographic variables according to common rules, identifying and separating address components (such as house number, street name, apartment number, rural route number, and post office box number) from the keyed address field (Laplant 1989), and "standardizing" an individual's first name by comparing the keyed first name to a library of nicknames and converting nicknames observed in the data to their common antecedent names (Paletz 1989).

The census/PES record-linkage procedure is a weight-based procedure. The determination of a weighting method includes consideration of both model-based and *ad hoc* rules for assigning weights for agreement and disagreement on individual fields of information, rules for assigning weights for close but inexact agreement on particular fields, rules for assigning weights when information is missing from records, and rules for assigning weights when certain combinations of variables are found to be in agreement or disagreement.

The designation of candidate matched pairs in census/PES matching reflects certain constraints that are placed on the matching process. First, time and resource constraints make it impractical to compare each record in one data file to every record in the other data file. Accordingly, comparisons are made only between pairs of records that meet certain minimal criteria, such as that they fall in the same census block and share the same first letter of last name. The subset of records formed by this restriction is referred to as a "block", and the variables required to be in agreement for a match to be declared are referred to as "blocking variables" (Jaro 1989).

Another constraint placed on the census/PES matching operation is that a given record in one data file is not allowed to be declared matched to more than one record in the other data file. The approach that is used to perform the assignment of candidate matched pairs draws on operations-research techniques for solving the so-called transportation problem (Jaro 1989). The algorithm assigns candidate matches so as to maximize the sum of composite weights among all possible pairs of records within a block defined by the blocking variables, subject to the aforementioned restriction that no record is allowed to match more than one record in the other data file. For example, suppose that within a particular block record A from file 1 has a higher agreement weight with record B from file 2 than with any other record in file 2. The assignment algorithm still might link record A to another record, say C, and link B to another record, say D, if the sum of the agreement weights for (A,C) and (D,B) are higher than for other permutations of candidate match assignment.

The current approach to census/PES matching contemplates three possible actions to be taken by the computer: declare a record pair to be a match, declare a record pair to be a "possible match", or declare a record to be not matched. All non-matches and possible matches are sent to clerks to be reviewed, and an attempt to obtain a followup interview is made for households where there is a discrepancy between the census and the PES. The distinction between possible matches and non-matches only has to do with the procedures applied by clerks when they review these cases (Childers 1989; Donoghue 1990). In the processing of 1990 census/PES data, the operator of the matching program set cutoff weights manually to distinguish matches, possible matches, and non-matches after scanning sets of candidate matched pairs with weights in a certain range. A new technique by Belin and Rubin (1991) offers an alternative for automating the setting of cutoffs.

3. AN EXPERIMENT

3.1 Factors Influencing the Output of Record-Linkage Procedures

The performance of a record-linkage procedure can depend on a number of factors, including:

- (1) The choice of matching variables;
- (2) The choice of blocking variables;
- (3) The assignment of weights to agreement or disagreement on various matching variables;
- (4) The handling of close but not exact agreement between matching variables;
- (5) The handling of missing data in one or both of a pair of records;
- (6) The algorithm for assigning candidate matches;
- (7) The choice of a cutoff weight above which record pairs will be declared matched;
- (8) The site or setting from which the data are obtained.

Among these factors, only (8) represents a source of variation over which the operator of the matching program does not have control. As mentioned earlier, two lines of inquiry are of primary interest in the experiment. Identifying major sources of variability in record linkage could help to focus future record-linkage research and to offer a deeper understanding of the process that generates errors in linkage procedures. Further, it is of interest to identify the combination of factors that works best in achieving a maximum number of matches while maintaining low error rates, since in practice the user generally must make a single choice among a myriad of possibilities for each factor just described.

3.2 Factorial Experiment Using Census/ Post-Enumeration Survey Data

A study was conducted using data from each of the three sites (St. Louis, Missouri; East Central Missouri including the Columbia, Missouri area; and a rural area in eastern Washington state) of the 1988 dress rehearsal census and PES. These data sets had been matched by computer and then reviewed by clerks. For the purposes of subsequent analysis, the final clerical determinations of true and false match status are taken as the truth. Thus, although subsequent analyses will only be as accurate as the determinations by clerks, these data files offer an excellent opportunity to study record linkage.

Descriptions of the specific methods used in linking records between the census and PES can be found in Jaro (1989), Winkler (1991), and Winkler and Thibaudeau (1992). The current implementation of the record-linkage procedure allows the user a variety of options over all of the factors listed in Section 3.1 except for the choice of an algorithm for assigning candidate matches (a “linear-sum assignment” algorithm is used; see Jaro 1989).

The variables available for matching census/PES records include name, address, age, race, sex, telephone number, marital status, and relationship to head of household. In practice, name is usually broken down into first name, last name, and middle initial, with these three used as separate matching variables. A preprocessing

program is typically used to parse address information into house number, street name, apartment number, rural route number, and box number (Laplant 1989). Sometimes “irregularities” in address information, perhaps caused by clerical typing errors or by recording errors on the part of a census or post-enumeration survey interviewer, result in an inability to parse an address into various components; in these cases, the entire address field (referred to as the “conglomerated address”) is used as a matching variable. An available preprocessing program also can be used to convert nicknames to a “standardized” name using a library of names and their common variants (Paletz 1989). A variety of schemes are available for assigning weights based on close agreement between variables, and a procedure is also available for adding or subtracting weight to the composite weight for a record pair when certain combinations of fields are in agreement or disagreement (Winkler 1991).

The experiment consisted of eight “treatment” factors and one “blocking” factor (where “blocking” here refers to the experimental-design notion of a grouping of units expected to yield results as similar as possible in the absence of treatment effects) with replication across three sites in a $2^5 \times 3^3 \times 5 \times 13$ factorial design. The outcome variable in the experiment, described further in Section 3.5, was a transformation of the false-match rate, where the transformation was used to stabilize the variance of the outcome. The factors in the experiment can be described as follows:

Label	Description of factor	Number of levels of factors	Description of levels of factor
A	Assignment of weight for name fields.	5	<ol style="list-style-type: none"> 1. Assign weights of ± 2 for agreement/disagreement on first, last name. 2. Assign weights of ± 4 for agreement/disagreement on first, last name. 3. Assign weights of ± 6 for agreement/disagreement on first, last name. 4. Assign weights based on estimates of probabilities of agreement on first, last name from Fellegi-Sunter algorithm (see Winkler and Thibaudeau 1992). 5. Use frequency-based weighting for first, last name (see Winkler and Thibaudeau 1992).

Label	Description of factor	Number of levels of factors	Description of levels of factor
B	Assignment of weight for close but inexact agreement on name fields.	3	<ol style="list-style-type: none"> 1. Assign disagreement weight for any discrepancy in first, last name. 2. Assign fraction of agreement weight for close agreement on first, last name using Jaro string comparison metric (Jaro 1989; Winkler 1991). 3. Assign fraction of agreement weight for close agreement on first, last name using piecewise linear metric described in Winkler (1991).
C	Assignment of weight for non-name fields.	2	<ol style="list-style-type: none"> 1. Assign weights of ± 2 for agreement/disagreement on age, phone number, and address fields, and assign weights of ± 1 for agreement/disagreement on sex, race, marital status, relationship to head of household, middle initial. 2. Assign weights based on estimates of probabilities of agreement from Fellegi-Sunter algorithm.
D	Assignment of weight for close but inexact agreement on non-name fields.	3	<ol style="list-style-type: none"> 1. Assign disagreement weight for any discrepancy in non-name fields. 2. Assign fraction of agreement weight for close agreement on house number, street name, phone number, age, using Jaro string comparator. 3. Assign fraction of agreement weight for close agreement on street name using Jaro string comparator, for age using Jaro pro-rated-to-absolute-difference metric, for house number and phone number using Winkler piecewise-linear string comparator.
E	Use of keyed first name or standardized version of first name.	2	<ol style="list-style-type: none"> 1. Use the version of the individual's first name that was keyed into each data file for comparison of first name. 2. Use the version of the individual's first name that is obtained as output from name standardization software (Paletz 1989).

Label	Description of factor	Number of levels of factors	Description of levels of factor
F	Adjustment of weights for correlated agreement.	2	<ol style="list-style-type: none"> 1. Do not adjust the composite weight for possible correlated agreement. 2. Adjust composite weights for possible correlated agreement between first name, middle initial and among first name, sex, age.
G	Inclusion of marital status, relationship to head of household as matching variables.	2	<ol style="list-style-type: none"> 1. Do not include marital status, relationship as matching variables. 2. Include marital status, relationship as matching variables.
H	Use of four or seven digits of phone number.	2	<ol style="list-style-type: none"> 1. Use only last four digits of phone number as a matching variable. 2. Use all seven digits of phone number.
I	Site of census/post-enumeration survey.	3	<ol style="list-style-type: none"> 1. Eastern Washington state. 2. Columbia, Missouri. 3. St. Louis, Missouri.
J	Proportion of PES file declared matched.	13	1.-13. Let the number of records accepted as declared matches equal 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, 90% of the number of PES records in the given site.

With reference to the sources of variation described in Section 3.1, factors E, G, and H relate to the choice of matching variables; factors A, C, and F relate to the choice of a weighting scheme; factors B and D relate to the handling of close but inexact agreement; factor J reflects the choice of a cutoff; and factor I reflects the influence of the particular site on the performance of the matching procedure.

Consideration of resource limitations led to a decision not to address the effect of varying missing data treatments or the effect of different choices of blocking variables in this experiment, and the lack of available software precluded any investigation of alternative algorithms for assigning candidate matches. Belin (1989a, 1989b) studied the influence of missing data treatments and of different choices of blocking variables in an experiment similar to the factorial experiment described here. The results of that investigation suggested that alternative treatments of missing data had no substantial effect on false-match rates

associated with different cutoffs in matching of census/PES data, but the choice of blocking variables did have a substantial effect.

In this investigation, as in Belin (1989a, 1989b), only “one-pass” matching procedures are considered. That is, the entire computer-matching operation consists of a single cycle of choosing blocking variables, establishing weights, and setting a cutoff, as opposed to “multiple-pass” procedures that first use very restrictive blocking variables to skim off the nearly perfect matches, then relax the blocking criteria in successive passes through the data. The author is aware of very little research on multiple-pass matching procedures. Belin (1989b) reports that when single-pass procedures are used, procedures that use relatively less restrictive blocking criteria enjoy advantages over procedures that use relatively more restrictive blocking criteria, confirming the intuitive notion that the blocking process can exclude true matches from consideration as an unfortunate side effect.

3.3 Subtleties in Experimental Treatments

3.3.1 Treatments for Assigning Weights for Agreement/Disagreement on Fields of Information

To clarify the experiment, we describe each of the experimental factors in greater detail. Factors A and C are concerned with the assignment of weights for agreement and disagreement on the various matching variables. The different weighting approaches used in factors A and C include completely *ad hoc* methods and methods that are based on estimates of parameters in explicit probability models. The study of *ad hoc* weights provides an opportunity to gauge the importance of incorporating more complicated approaches to weighting.

The *ad hoc* weighting schemes call for a weight of U , say, to be added to the composite weight if the fields being compared agree, and for an identical weight U to be subtracted from the composite weight if the fields being compared disagree. Three different values of U are studied in factor A, with the same value of U being assigned for agreement on first name as for agreement on last name. In factor C, an *ad hoc* scheme that weights some variables more than others is studied, with the decision about which variables to weight more being based on *a priori* judgments. Belin (1989b) suggests that such a “modified-equal-weighting” scheme has advantages over an “equal-weighting” scheme in which all matching variables are assigned the same weights for agreement or disagreement.

The “Fellegi-Sunter algorithm” refers to the method outlined in Fellegi and Sunter (1969), which is based on a probabilistic model that incorporates information about patterns of agreement and disagreement between pairs of records. The model postulates that probabilities of agreement on individual fields of information given that a pair is a true match are independent across all fields of information, and that independence across fields also holds given that a pair is a false match. The paper by Fellegi and Sunter shows that such a model implies certain optimality properties for the type of weighting scheme used by Newcombe *et al.* (1959), in which weights for individual fields of information are calculated by taking the logarithm of the ratio of probability of agreement given true match to the probability of agreement given false match, and in which composite weights are obtained by summing individual field weights.

In applications, the probabilities of agreement given true match and agreement given false match need to be estimated. For the treatments in the experiment characterized as relying on the Fellegi-Sunter weighting approach, the probabilities of agreement given true match are estimated using a version of an EM algorithm (Dempster, Laird and Rubin 1977) to obtain maximum likelihood estimates of these probabilities based on counts of all possible patterns of agreement observed in the data files at hand (Winkler 1989; Jaro 1989). The probabilities of agreement given

false match are estimated based on counts of agreement on individual fields between all record pairs that agree on blocking variables, making use of the fact that most of the pairs that could possibly be brought together as matches are not true matches (Winkler and Thibaudeau 1992).

Another weighting approach that has been implemented in the Census Bureau’s record linkage software considers the relative frequency of names in the data files at hand, assigning more weight for agreement on names such as Abramowicz, which may be relatively rare, than for agreement on names such as Smith, which may be common. Of course, it could happen that in a particular area Abramowicz is a more common name than Smith, in which case the frequency-based weighting approach would assign greater weight to agreement on the name Smith. The idea of incorporating information on marginal frequencies from the current data files was mentioned by Newcombe *et al.* (1959), and has been noted by many authors since then, including Fellegi and Sunter (1969). (Thus, the distinction drawn here between the “Fellegi-Sunter algorithm” and “frequency-based weighting” is actually a distinction between two methods of calculating weights that are both discussed by Fellegi and Sunter.) Details on the implementation of frequency-based weighting in the Census Bureau’s software can be found in Winkler and Thibaudeau (1992).

3.3.2 Treatments for Handling Close but Inexact Agreement

Factors B and D deal with the handling of fields that may agree closely but do not agree exactly with one another. Several techniques have been proposed for handling close but inexact agreement between fields of information, often reflecting different perspectives on probable departures from exact agreement.

The Jaro string comparator is designed to measure the closeness of agreement of two multi-character fields; the metric that defines closeness is a function of the lengths of the character fields in the two files, the number of characters in common between the character fields, and the number of transpositions of characters between the character fields. The weight that gets assigned for partial agreement is between the weight for agreement on the field and the weight for disagreement on the field, and is a linear function of the string comparator metric between the agreement weight and the disagreement weight.

The Winkler piecewise-linear approach uses the same metric as the Jaro string comparator to define closeness of agreement, but the rate at which partial agreement weights decrease from the agreement weight to the disagreement weight is a piecewise linear function of the string comparator metric, requiring two user-supplied rate parameters and two user-supplied thresholds where the slope changes.

The Jaro pro-rated method assigns a weight between the agreement weight and the disagreement weight based on the absolute value of the difference between two numeric fields. As with the aforementioned techniques, the partial agreement weight falls off as a linear function of the absolute value of the difference.

Even for some numeric fields (e.g., telephone number), a comparison method designed to accommodate slight typographical variation would seem more sensible than a method based on absolute numerical difference. However, for variables such as year of birth or age, it may not be clear whether to target efforts toward accommodating typographical errors (for which a string comparison method would be best suited), reporting errors (for which the absolute-difference method may be most appropriate), or other types of errors such as “heaping” or rounding of reported ages on multiples of five years (for which neither of the previously mentioned comparison methods would be ideally suited). Accordingly, we pursue our empirical evaluations in an attempt to shed light on these issues.

3.3.3 Treatments Involving the Choice of Matching Variables

As mentioned previously, an approach has been developed at the Census Bureau for converting nicknames to a standardized root. Software developed by Paletz (1989) implements the name-standardization routine.

The treatment that omits marital status and relationship to head of household as matching variables allows for an assessment of the importance of two background demographic variables on the quality of matching. Chernoff (1980) develops theory for the information carried by a matching variable and shows that a variable recorded in error even a small percentage of the time can lose a substantial amount of information for matching purposes (e.g., the Kullback-Leibler information associated with a binary variable recorded in error three percent of the time is only about half that of a binary variable recorded without error). Considering that relationship to head of household could differ between the census and PES if the person listed as the head of household is different, and that marital status will change for some individuals in the intervening time, it is not clear in advance how much information for matching is provided by these variables. On the other hand, it is hard to imagine that using additional matching variables would be deleterious, so that this treatment provides a standard for assessing the practical significance of some of the other treatments.

The treatment of using either four or seven digits of phone number as a matching variable is self-explanatory. A motivation for considering this treatment is that one of the specific piecewise-linear string comparator methods proposed by Winkler was developed based on analysis of the last four digits of phone number as a matching variable.

3.3.4 Treatment for Adjusting Composite Weights for Correlated Agreement

The method described as adjusting the composite weight to reflect the possibility of correlated agreement is also due to Winkler and is described in Winkler and Thibaudeau (1992). Research by Kelley (1986) and Thibaudeau (1989) reveals that agreement on the various fields available for matching between the census and PES data files is far from being independent across fields. In particular, analyses suggested that agreement on first name was correlated with agreement on middle initial and that agreement on first name, age, and sex were mutually correlated. These findings led to the implementation of modifications to the composite weight when certain patterns appear (e.g., if first name, age, and sex all disagree, then a large value is subtracted from the composite weight). The current scheme for adjusting the composite weight is entirely *ad hoc*; research into methods that reflect correlated agreement still appears to be in its infancy.

3.4 Data Files Used in Experiment

As mentioned before, the three sites of the 1988 dress rehearsal census and post-enumeration survey provided separate data files on which these analyses of record linkage could be performed. There were 12,072 records in the PES file from St. Louis, 6,581 records in the PES file from East Central Missouri, and 2,782 records in the PES file from eastern Washington state. As was also noted earlier, the final determinations by clerks who reviewed these files were taken as the truth for purposes of evaluation. Other test censuses were conducted during the 1980's; the primary reason for not including the data from other test censuses in this experiment is that a considerable amount of “overhead” time is required to prepare a data set for the analyses performed here.

3.5 Outcome Variable

The primary outcome variable considered in this experiment was a transformation of the false-match rate. The false-match rate is defined as the number of false matches divided by number of declared matches, and is a common measure of performance in the literature on record linkage (e.g., Fellegi and Sunter (1969) attempt to provide output that satisfies a fixed false-match rate criterion supplied by the operator of the program). In order to stabilize the variance of the outcome, the analyses here use the arcsine of the square root of the false-match rate as an outcome variable.

3.6 Choice of Cutoff Weight as a Blocking Factor

It is clear that the false-match rate in record linkage is apt to depend heavily on the choice of a cutoff between declared matches and declared non-matches. Accordingly,

a blocking factor (Factor J) is introduced to fix the determination of cutoffs so as to facilitate comparison of other record-linkage treatments. To provide a standard for comparisons across sites having different numbers of records, the cutoff level is defined in terms of the proportion of the PES data file declared matched.

Because of the discreteness of record-linkage weights, it is possible to have ties among the weights of record pairs on the boundary where the cutoff should be assigned. For example, in a file of 10,000 records, there may be 40 records with weight W (of which 10 may be false matches), 7,980 records with weight greater than W (of which 3 may be false matches), and 1,980 records with weight less than W . If the treatment in factor J calls for 80% of the PES file to be matched, then it may not be obvious how to calculate the false-match rate, since there are 40 records with the same weight straddling the point where the cutoff should be set. Calculations of the false-match rate in such a case are based on the following relationship:

$$\text{fmr} = \frac{f_{\text{abv}} + \frac{f_{\text{bdy}}}{n_{\text{bdy}}} \times (n_{\text{cut}} - n_{\text{abv}})}{n_{\text{cut}}},$$

where fmr denotes false-match rate, f_{abv} is the number of false matches and n_{abv} the number of declared matches with weights above the cutoff weight, f_{bdy} is the number of false matches and n_{bdy} the number of declared matches with weights equal to the boundary cutoff weight, and n_{cut} is the number of declared matches needed to satisfy the condition that a certain percentage of the PES data file be declared matched. If we were to calculate the false-match rate by randomly selecting the appropriate number of boundary records to satisfy the cutoff criterion, then the expression above would give the expected false-match rate over repetitions of such a procedure; thus, the logic behind this definition is clear.

In the example above, one fourth of the boundary cases are false matches, and twenty additional records are needed to satisfy the stipulation that 80% of the file be declared matched. Effectively five false matches are added to the three among the records among the pairs with weights above the cutoff weight, giving a false-match rate of $(3 + 0.25(40 - 20))/8,000 = 8/8,000 = 0.001$.

3.7 Further Considerations Relevant to the Analysis of Experimental Results

Analysis of the experimental results proceeded from the standpoint that general indications of significance are more important than precise p -values, especially because the experiment itself is exploratory. Belin (1991) points out that appropriate methods for assessing significance from these data are somewhat complicated; this is because site

should be thought of as a random factor (since we would like to generalize about treatment effects from the sample of three sites to a population of many possible sites), but standard procedures that use the site by treatment interaction as the error term for a particular treatment suffer from low power given the small number of available sites. Belin (1991) uses the Johnson-Tukey display-ratio plot (Johnson and Tukey 1987), which is a close relative of the half-normal plot of Daniel (1959), to estimate underlying noise levels in assessing the significance of effects. In this paper, we do not attempt to present formal significance findings.

4. RESULTS

4.1 ANOVA Breakdown of Experimental Results

We begin by breaking down the results of the factorial experiment into an analysis of variance, distinguishing treatment effects, site effects, cutoff effects, and their interactions from one another, grouping effects of the same order. Table 4.1 is an excerpt from the complete ANOVA breakdown of the experiment, showing treatment interactions up to four-way along with corresponding error terms.

F -statistics are calculated dividing the mean square for the given effect by the mean square for the effect-by-site interaction term. Thus, for example, the F -statistic for three-way interactions among treatments is calculated as $0.0120/0.00470 = 2.551$, with the denominator coming from the line for the four-way treatment-by-site interaction.

If the F -statistics are interpreted in the usual way, then statistical significance at the 0.0001-level is achieved for all of the F -statistics reported in Table 4.1 except the treatment-by cutoff four-way interactions; however, caution should be used in interpreting these results. First, the magnitudes of the various mean-square terms suggest that the higher-order effects are not of substantial practical importance. Further, the comparison of the F -statistics calculated above to a reference F -distribution relies on certain exchangeability assumptions (*e.g.*, that site-to-site variability in main effects is the same for all main effects) that are not necessarily well-founded. For example, it may not make sense to pool site-to-site variability in the effect of four versus seven digits of phone number with site-to-site variability in the effect of the different weighting schemes in estimating an error term for main effects.

4.2 Importance of Choice of Cutoff as Compared to Other Controllable Factors

It is evident (*e.g.*, from the mean squares for main effects) that site-to-site variability and variability due to the choice of a cutoff are considerably larger than the variability explained by differences in treatments. Although

Table 4.1

Excerpt from ANOVA Breakdown of Factorial Experiment, Grouping Effects of the Same Order

Source	<i>df</i>	Sums of squares	Mean square	<i>F</i>
Site main effects	2	35.195	17.598	
Treatment main effects	13	30.917	2.378	10.570
Cutoff main effects	12	147.515	12.293	7.548
Treatment/site 2-way interactions	26	5.850	0.225	
Cutoff/site 2-way interactions	24	39.089	1.629	
Treatment/treatment 2-way ints	70	6.992	0.100	4.041
Treatment/cutoff 2-way ints	156	1.410	0.009	3.553
Treatment/site 3-way interactions	140	3.461	0.0247	
Cutoff/treatment/site 3-way ints	312	0.794	0.0025	
Treatment 3-way interactions	206	2.472	0.0120	2.551
Treatment/cutoff 3-way ints	840	0.530	0.0006	1.866
Treatment/site 4-way interactions	412	1.938	0.00470	
Cutoff/treatment/site 4-way ints	1,680	0.568	0.00034	
Treatment 4-way interactions	365	0.747	0.00205	2.365
Treatment/cutoff 4-way ints	2,472	0.267	0.00011	0.236
Treatment/site 5-way interactions	730	0.632	0.00087	
Cutoff/treatment/site 5-way ints	4,944	0.226	0.00046	
...				
Total	56,159	279.169		

this result may be explained in part by the fact that some treatments are very close to one another (*e.g.*, using four digits versus seven digits of phone number), it is nevertheless the case that some of the qualitative differences between treatments are quite substantial (*e.g.*, leaving out two matching variables versus keeping them in). The ANOVA breakdown also highlights the fact that we can expect substantial site-to-site variability in false-match rates. In their approach to calibrating record-linkage procedures, Belin (1991) and Belin and Rubin (1991)

explicitly accommodate site-to-site variability in providing estimates of false-match rates corresponding to different cutoffs.

4.3 The Main Effects of Treatments

In Table 4.2, we give the mean of the outcome variable observed for each level of the treatment factors. Since arcsine (x) is a monotone increasing function of x , lower values of the outcome signify lower false-match rates and thus better performance.

Table 4.2

Marginal Values of arcsine($\sqrt{\text{fmr}}$) for each Level of Experimental Treatments Averaged over all other Experimental Conditions

Factor	A	(name wts)	Factor	B	(inexact agree, name wts)	Factor	C	(non-name wts)
Level	1	0.106	Level	1	0.113	Level	1	0.101
	2	0.096		2	0.094		2	0.101
	3	0.093		3	0.095			
	4	0.130						
	5	0.079						
Factor	D	(inexact agree, non-name wts)	Factor	E	(Standardize name)	Factor	F	(Adjust for correlated agree)
Level	1	0.111	Level	1	0.102	Level	1	0.106
	2	0.108		2	0.100		2	0.095
	3	0.084						
Factor	G	(Include marit/rel)	Factor	H	(Four or seven digits phone #)			
Level	1	0.103	Level	1	0.102			
	2	0.098		2	0.100			

Belin (1991) breaks down the experimental findings into a set of complementary orthogonal contrasts. The largest main-effect contrasts among those prespecified by Belin (1991) were those between frequency name weights ($A = 5$) and Fellegi-Sunter name weights ($A = 4$), between Winkler's string comparators on non-name fields ($D = 3$) and Jaro's corresponding string comparators ($D = 2$), between some string comparator for names ($B = 2$ or 3) and no string comparator for names ($B = 1$), between some string comparator for non-name fields ($D = 2$ or 3) and no string comparator for these fields ($D = 1$), and between performing an adjustment for correlated agreement ($F = 2$) and not performing such an adjustment ($F = 1$).

4.4 Two-Way Treatment Interactions

The largest two-way treatment interaction contrast among those reviewed by Belin (1991) was the $F \times G$ effect, which is the interaction of performing an adjustment for correlated agreement (among first name and middle initial and among first name, age, and sex) with including or not including marital status and relationship to head of household as matching variables. This contrast was statistically significant according to any of the procedures used in Belin (1991) for estimating a background noise level. We show the average levels of the outcome across the four treatment combinations above in Table 4.3.

Table 4.3
Average Performance for Combinations of
F and G Treatments

F	G	False-match rate	Arcsine($\sqrt{\text{fmr}}$)
1	1	0.0182	0.116
1	2	0.0143	0.097
2	1	0.0128	0.091
2	2	0.0151	0.100

This result suggests that the adjustment for correlated agreement (level 2 of factor F) helps a great deal when marital status and relationship are not included as matching variables (level 1 of factor G), but the adjustment for correlated agreement does not help on average when marital status and relationship are included as matching variables. That we are able to identify this type of effect emphasizes the importance of pursuing empirical evaluations in an experimental framework.

The next two largest two-way treatment interaction contrasts cited by Belin (1991) after the $F \times G$ interaction comprise part of the $A \times B$ interaction (involving the choice of name weights and the choice of string comparisons to use for name fields). We show the average results for all of the combinations of treatments for factors A and B below as Table 4.4.

Table 4.4
Average Performance for Combinations of
A and B Treatments

A	B	False-match rate	Arcsine ($\sqrt{\text{fmr}}$)
1	1	0.0192	0.120
1	2	0.0140	0.099
1	3	0.0143	0.100
2	1	0.0170	0.110
2	2	0.0120	0.087
2	3	0.0123	0.089
3	1	0.0177	0.113
3	2	0.0118	0.084
3	3	0.0119	0.083
4	1	0.0254	0.145
4	2	0.0193	0.123
4	3	0.0189	0.122
5	1	0.0109	0.079
5	2	0.0109	0.079
5	3	0.0109	0.078

Thus, we find that when we use frequency-based name weights ($A = 5$), it hardly matters whether we use any string comparison method, but when we use *ad hoc* name weights or Fellegi-Sunter name weights, the use of string comparison methods substantially improves the average performance of the computer-matching procedure.

We highlight some of the other interesting findings noted in Belin (1991) based on exploring the largest two-way treatment interaction effects:

- (1) The Winkler approach to inexact agreement on non-name variables (*i.e.*, $D = 3$), which is the best treatment on average for factor D, has more of a helpful effect on average when marital status and relationship to head of household are included as matching variables (*i.e.*, $G = 2$), even though the latter variables are not included in any of the treatments for handling inexact agreement.
- (2) Unlike the other treatments for name weights, which appear to be helped by the inclusion of marital status and relationship, frequency-based name weighting appears to be adversely affected by the inclusion of these variables.
- (3) *Ad hoc* weights of ± 6 for agreement on name perform better on average when combined with the *ad hoc* weighting approach to non-name variables; *ad hoc* name weights of ± 4 and ± 2 work better with the weights assigned by the Fellegi-Sunter algorithm to non-name variables.
- (4) Without the adjustment for correlated agreement, Fellegi-Sunter weights for non-name variables worked better for these data than *ad hoc* weights, but the *ad hoc* weights worked better when the adjustment for correlated agreement was included. (However, based on the method of estimating the background noise level described in Belin (1991), this phenomenon should not necessarily be expected to carry over to other sites.)

4.5 Which Treatment Combination Works Best?

To wrap up the analysis of the experimental results, we consider now the question of which treatment combination works best. To measure the performance for a given treatment combination, we take the average outcome from using that procedure across the three available sites. The outcomes we examine are the false-match rates corresponding to 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, and 90% of the PES file declared matched. The results from the experiment are provided in Table 4.5.

Table 4.5

Best Treatment Combination for each of Thirteen Cutoffs from Factorial Experiment

Cutoff level	Levels of factors in best treatment combination (A B C D E F G H)								False-match rate for best treatment combination averaged over three sites
60% matched	3	3	2	3	2	1	1	1	0.00042
62.5% matched	3	3	1	3	1	1	2	1	0.00047
65% matched	3	3	1	3	2	2	2	2	0.00052
67.5% matched	3	3	2	3	2	2	1	1	0.00071
70% matched	2	3	2	3	1	2	1	2	0.00079
72.5% matched	5	2	2	3	1	1	1	2	0.00081
75% matched	5	1	1	3	2	1	2	1	0.00112
77.5% matched	3	3	1	3	2	1	2	1	0.00133
80% matched	2	3	2	3	1	2	1	2	0.00188
82.5% matched	3	3	1	3	2	2	1	1	0.00571
85% matched	5	1	2	3	2	2	1	2	0.01556
87.5% matched	2	3	2	3	1	2	1	2	0.03023
90% matched	2	3	2	3	1	2	1	2	0.05174

These results contrast with the earlier result suggesting that frequency-based weighting for names (level 5 for factor A) is better on average than using *ad hoc* name weights of ± 6 (level 3 for factor A). Apparently, the reason that the latter is worse on average is due to certain interaction effects. When the *ad hoc* weighting approach is combined with the appropriate levels of other factors, it appears to perform at least as well as the frequency-weighting approach. We also note that the best combination of factors F and G is not always treatments 2 and 1, respectively, despite our earlier finding that this treatment combination for these two factors performs best on average. Only treatment 3 of factor D (using Winkler modifications in handling inexact agreement on non-name variables) is an unequivocal choice for the best treatment no matter how we measure the outcome of the experiment. The choice for the best treatment for name weights is between deterministic weights of ± 6 or ± 4 and the frequency name-weighting approach. If one of the deterministic weighting schemes is used, the Winkler approach

to string comparisons for names is to be recommended; with frequency name weights, it is not clear that any string comparison approach should be used on names.

Between Fellegi-Sunter weights for non-name variables and *ad hoc* weights, the choice is not obvious, but earlier analysis suggested that the effect either way is small. Similar remarks apply to the choice of whether to use standardized or unstandardized first names and to the choice of whether to use four or seven digits of the phone number.

Considering the fact that there is not a single treatment combination that is uniformly superior to all other treatment combinations, one might look to the performance of different treatment combinations in a particular region of interest (*e.g.*, where the false-match rate is around 0.001). However, if we look at the best treatment combinations in the region where 70%-80% of the PES file is declared matched (*i.e.*, restricting attention to five cutoffs), we still find no obvious choice for a preferred treatment combination. Averaged across those five cutoffs, the best treatment combination is (2,3,2,3,1,2,1,2); that is, using name weights of ± 4 , incorporating Winkler's modifications to inexact agreement on name, estimating weights using the Fellegi-Sunter algorithm for non-name variables, using Winkler's approach to inexact agreement for non-name variables, using the original unstandardized version of first name, adjusting the composite weight for correlated agreement, not including marital status and relationship to head of household as matching variables, and using all seven digits of phone number.

For comparison, we display in Table 4.6 the average performance of some of the other candidates for best treatment combination. Thus it appears that the best alternatives to (2,3,2,3,1,2,1,2) are treatment combinations (3,3,1,3,2,2,2,2) and (3,3,1,3,2,1,2,1). Both of these procedures feature name weights of ± 6 , predetermined

Table 4.6

Average False-match Rates for Different Treatment Combinations Across Three Sites and across Five Cutoff Levels (70%, 72.5%, 75%, 77.5%, and 80% of PES File Declared Matched)

Levels of factors in treatment combination (A B C D E F G H)	Average false-match rate across sites and across cutoffs with 70%, 72.5%, 75%, 77.5%, and 80% of PES file declared matched
3 3 2 3 2 1 1 1	0.00493
3 3 1 3 1 1 2 1	0.00154
3 3 1 3 2 2 2 2	0.00137
3 3 2 3 2 2 1 1	0.00161
2 3 2 3 1 2 1 2	0.00124
5 2 2 3 1 1 1 2	0.00191
5 1 1 3 2 1 2 1	0.00153
3 3 1 3 2 1 2 1	0.00138
3 3 1 3 2 2 1 1	0.00156
5 1 2 3 2 2 1 2	0.00155

ad hoc weights for non-name variables, Winkler's approaches to inexact agreement for both name and non-name variables, standardized first names, and inclusion of marital status and relationship as matching variables. These treatment combinations differ from each other in that one includes an adjustment of the composite weight for correlated agreement and calls for using seven digits of phone number, whereas the other features no adjustment of weights for correlated agreement and only four digits of phone number. The treatment combinations involving the use of frequency-based name weighting do not perform as well as the best treatment combinations using *ad hoc* name weights according to this standard.

In the 1990 PES, the treatment combination that was used in computer-matching operations was very close to treatment combination (5,3,2,3,2,2,2,1). In the test-census data sets studied here, this treatment combination produced an average false-match rate across the five cutoffs of 0.00179.

4.6 Concluding Remarks

While the results in this paper address the tradeoff between the number of records declared matched and false-match rates, an anonymous referee noted that "every gain which is achieved by a superior record linkage procedure must be justified by the cost of implementing that procedure." This is another tradeoff that any practitioner can appreciate. Hopefully, the findings presented here about the relative importance of various factors in record linkage will provide some guidance to those who develop and implement linkage software. Because some of the results may depend on specific features of the census/PES data being matched, there may be some question as to how these results relate to other record-linkage settings. But as was emphasized at the outset, one practical recommendation that does generalize across data settings is the call for taking an experimental approach to the study of record linkage. Empirical study through designed experiments is a tried and true source of guidance, offering a clear framework for adding to the accumulated insights of record-linkage specialists.

ACKNOWLEDGMENTS

Much of this work was done while the author was working for the Record Linkage Staff of the U.S. Bureau of the Census in Washington, D.C. The author gratefully acknowledges helpful discussions and comments from Don Rubin, Bill Winkler, Alan Zaslavsky, and an anonymous referee, as well as earlier support from JSA 88-02 and JSA 89-07 while the author was a doctoral candidate at Harvard University.

REFERENCES

- ABBATT, J.D. (1986). A cohort study of eldorado uranium workers. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 51-57.
- ACHESON, E.D. (1967). *Medical Record Linkage*. Oxford: Oxford University Press.
- ACHESON, E.D. (Ed.) (1968). *Record Linkage in Medicine*, Edinburgh: E. & S. Livingstone.
- BALDWIN, J.A., ACHESON, E.D., and GRAHAM, W.J. (Eds.) (1987). *A Textbook of Medical Record Linkage*. Oxford: Oxford University Press.
- BELIN, T.R. (1989a). Outline of procedure for evaluating computer matching in a factorial experiment. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- BELIN, T.R. (1989b). Results from evaluation of computer matching. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- BELIN, T.R. (1991). Using mixture models to calibrate error rates in record-linkage procedures, with application to computer-matching for census undercount estimation. Ph.D. thesis, Department of Statistics, Harvard University. (Published by University Microfilms, Inc.)
- BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.
- BOYLE, C.A., and DECOUFLÉ, P. (1990). National sources of vital status information: Extent of coverage and possible selectivity in reporting. *American Journal of Epidemiology*, 131, 160-168.
- BROWN, P., LAPLANT, W., LYNCH, M., ODELL, S., THIBAUDEAU, Y., and WINKLER, W. (1988). Collective Documentation for the 1988 PES Computer Match Processing and Printing. Vols. I-III, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- BUREAU OF THE CENSUS (1988-1991). 1990 Decennial Census Information Memorandum Series, Decennial Planning Division, Bureau of the Census, Washington, D.C.
- [Note: To all of the reports in the aforementioned memorandum series, the following statement is attached:
- "These overviews are prepared for use by planning and operating divisions within the Census Bureau who are conversant with the background, previous experiences, terminology, and processes, as well as with the overall framework of the decennial census design, goals, and inter-relationships of operations and systems. They are NOT [emphasis in original] intended or appropriate for external distribution and should not be sent outside the Census Bureau without prior approval from Jim Dinwiddie ([301]-763-5270) of the Decennial Planning Division."]
- BUREAU OF THE CENSUS (1987-1991). STSD Decennial Census Memorandum Series, Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.

- CARPENTER, M., and FAIR, M.E. (Eds.) (1990). Canadian Epidemiology Research Conference – 1989: *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario.
- CHERNOFF, H. (1980). The identification of an element of a Large population in the presence of noise. *Annals of Statistics*, 8, 1179-1197.
- CHILDERS, D. (1989). 1990 PES Within Block Matching – Clerical Matching Group. STSD Decennial Census Memorandum Series #V-69, U.S. Bureau of the Census, Washington, D.C.
- CITRO, C.F., and COHEN, M.L. (Eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, Washington, D.C.: National Academy Press.
- COHEN, M.L. (1990). Adjustment and reapportionment – Analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.
- COOMBS, J.W., and SINGH, M.P. (Eds.) (1988). *Proceedings of the Symposium on Statistical Uses of Administrative Data*. Statistics Canada, Ottawa, Ontario.
- COPAS, J., and HILTON, F. (1990). Record Linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A*, 153, 287-320.
- CURB, J.D., FORD, C.E., PRESSEL, S., PALMER, M., BABCOCK, C., and HAWKINS, C.M. (1985). Ascertainment of vital status through the National Death Index and the Social Security Administration. *American Journal of Epidemiology*, 121, 754-766.
- DANIEL, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1, 311-341.
- DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- DONOGHUE, G. (1990). Clerical Specifications for the 1990 Post Enumeration Survey Before Followup Matching – Special Matching Group. STSD Decennial Census Memorandum Series #V-92, U.S. Bureau of the Census, Washington, D.C.
- DULBERG, C.S., SPASOFF, R.A., and RAMAN, S. (1986). Reactor clean-up and bomb test exposure study. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 59-62.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and Beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.
- ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, 84, 927-944.
- FAGERLIND, I. (1975). *Formal Education and Adult Earnings: A Longitudinal Study on the Economic Benefits of Education*, Stockholm: Almqvist and Wiksell.
- FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 1-39.
- GOLDACRE, M.J. (1986). The Oxford record linkage study: Current position and future prospects. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press, 106-129.
- HILL, T. (1981). Generalized Iterative Record Linkage System: GIRLS. (Glossary, Concepts, Strategy Guide, User Guide), Systems Development Division, Statistics Canada, Ottawa, Ontario.
- HILL, T., and PRING-MILL, F. (1986). Generalized iterative record linkage system: GIRLS, (revised edition). Systems Development Division, Statistics Canada, Ottawa, Ontario.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, 46, 261-269.
- HOWE, G.R., and LINDSAY, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers in Biomedical Research*, 14, 327-340.
- HOWE, G.R., and SPASOFF, R.A. (Eds.) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press.
- JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.
- JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- JOHANSEN, H.L. (1986). Record linkage of national surveys: The Nutrition Canada example. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 153-163.
- JOHNSON, E.G., and TUKEY, J.W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao Data. In *Design, Data, and Analysis*, (Ed. C.L. Mallows) New York: John Wiley and Sons.
- JOHNSON, R.A. (1991). Methodology for Evaluating Errors in U.S. Department of Justice Attorney Workload Data. Unpublished technical report, General Accounting Office, Washington, D.C.
- KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.
- KERSHAW, D., and FAIR, J. (1979). *The New Jersey Income and Maintenance Experiment: Operations, Surveys, and Administration*, Volume I. New York: Academic Press.
- KILSS, B., and ALVEY, W. (Eds.) (1984a). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. I, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., and ALVEY, W. (Eds.) (1984b). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. II, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

- KILSS, B., and ALVEY, W. (Eds.) (1984c). *Statistics of Income and Related Administrative Record Research: 1984*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., and ALVEY, W. (Eds.) (1985). *Record Linkage Techniques - 1985*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., and ALVEY, W. (Eds.) (1987). *Statistics of Income and Related Administrative Record Research: 1986-1987*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., and JAMERSON, B. (Eds.) (1990). *Statistics of Income and Related Administrative Record Research 1988-1989*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., and SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin*, Vol. 41, 10, 14-22.
- LAPLANT, W. (1988). User's Guide for the Generalized Record Linkage Program Generator (GENLINK). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- LAPLANT, W. (1989). User's Guide for the Generalized Address Standardizer (GENSTAN). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.
- NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records (with discussion). *Journal of the American Statistical Association*, 87, 1193-1208.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., and ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado Uranium Workers. *Computers in Biology and Medicine*, 13, 157-169.
- NICHOLL, J.P. (1986). The use of hospital in-patient data in the analysis of the injuries sustained by road accident casualties. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 243-244.
- PALETZ, D. (1989). Name standardization software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- ROGOT, E., SORLIE, P.D., and JOHNSON, N.J. (1986). Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Disease*, 39, 719-734.
- ROGOT, E., SORLIE, P.D., JOHNSON, N.J., GLOVER, C.S., and TREASURE, D.W. (1988). A Mortality Study of One Million Persons. Public Health Service, National Institutes of Health, Washington, D.C.
- SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.
- SMITH, M.E., and NEWCOMBE, H.B. (1975). Methods for computer linkage of hospital admission-separation records for cumulative health histories. *Methods of Information in Medicine*, 14, 118-125.
- TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Section on Statistical Computing, American Statistical Association* 283-288.
- WENTWORTH, D.N., NEATON, J.D., and RASMUSSEN, W.L. (1983). An evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the ascertainment of vital status. *American Journal of Public Health*, 73, 1270-1274.
- WILLIAMS, B.C., DEMITRACK, L.B., and FRIES, B.E. (1992). The accuracy of the National Death Index when personal identifiers other than Social Security Number are used. *American Journal of Public Health*, 82, 1145-1147.
- WINKLER, W.E. (1985a). Preprocessing of lists and string comparison. In *Record Linkage Techniques - 1985*, (Eds. W. Alvey and B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 181-187.
- WINKLER, W.E. (1985b). Exact matching lists of businesses: blocking, subfield identification, and Information Theory. In *Record Linkage Techniques - 1985*, (Eds. W. Alvey and B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 227-241.
- WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter Model of record linkage. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 145-155.
- WINKLER, W.E. (1991). Documentation of record-linkage software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- WINKLER, W.E., and THIBAUDEAU, Y. (1992). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

The Discrimination Power of Dependency Structures in Record Linkage

YVES THIBAUDEAU¹

ABSTRACT

A record-linkage process brings together records from two files into pairs of two records, one from each file, for the purpose of comparison. Each record represents an individual. The status of the pair is a “matched pair” status if the two records in the pair represent the same individual. The status is an “unmatched pair” status if the two records do not represent the same individual. The record-linkage process is governed by an underlying probabilistic process. A record-linkage rule infers the status of each pair of records based on the value of the comparison. The pair is declared a “link” if the inferred status is that of a matched pair, and it is declared a “non-link” if the inferred status is that of an unmatched pair. The discrimination power of a record-linkage rule is the capacity of the rule to designate a maximum number of matched pairs as links, while keeping the rate of unmatched pairs designated as links to a minimum. In general, to construct a discriminatory record-linkage rule, some assumptions must be made on the structure of the underlying probabilistic process. In most of the existing literature, it is assumed that the underlying probabilistic process is an instance of the conditional independence latent class model. However, in many situations, this assumption is false. In fact, many underlying probabilistic processes do not exhibit key properties associated with conditional independence latent class models. The paper introduces more general models. In particular, latent class models with dependencies are studied and it is shown how they can improve the discrimination power of particular record-linkage rules.

KEY WORDS: Record-linkage rule; Latent class model; Expectation-Maximization procedures.

1. INTRODUCTION

The goal of the paper is to show how record-linkage rules can gain in discriminatory power when probabilistic models more descriptive of the underlying probabilistic processes, are elicited. For this purpose, a particular record-linkage situation is chosen and the conditional independence model, traditionally used in record linkage, is compared to a more descriptive model, in the sense that the new model allows for the expression of more complex relations of dependency between some of the variables involved.

First some terminology must be reviewed. In section 2, the definition of record-linkage process is stated and a general formulation of the probabilistic process underlying a record-linkage process is given. This formulation leads to the expression of two central concepts: the concepts of record-linkage rule and that of most discriminatory record-linkage rule.

In section 3, probabilistic models for record linkage are considered. In the first part of section 3, the family of latent class models is introduced and it is shown how this family provides natural models for the probabilistic process underlying a record-linkage process. In the second part of the section, the focus is on a particular model in the family of latent class models: the latent class model

with conditional independence. This model is of interest because it is easy to handle computationally. In the third part, inference techniques adapted to the conditional independence model are reviewed.

In section 4, an application is presented. For this application, truth and falsehood are available, that is, it is known which pairs are matched and which aren't. The first part describes how the information on truth and falsehood was obtained. The second part shows how dependencies between the comparison fields are generated. In the third part of section 4, the knowledge on truth and falsehood is used to evaluate the dependencies between the comparison fields. This leads in the fourth part to the formulation of a model more descriptive of the underlying probabilistic structure of the record-linkage process. The final part is a brief discussion regarding the techniques of parameter estimation for generalized latent class models.

In section 5, an alternative methodology to construct approximate probabilistic models is presented. The model produced by this methodology is compared to those introduced in sections 3 and 4, in terms of discrimination power of the record-linkage rules derived from the models. The results of the comparisons are reported in section 6. In section 7, the suggestions of an anonymous referee to improve the methodology of the paper are presented. In section 8, conclusions are drawn and guidelines are provided.

¹ Yves Thibaudeau, U.S. Bureau of the Census, Federal Bldg. 4, Room 3000, Washington, D.C. 20233.

2. THE FELLEGI-SUNTER MODEL FOR RECORD-LINKAGE

2.1 Record-Linkage Processes

The paper is geared toward building new record-linkage techniques. Before expanding on new record-linkage techniques, some background is necessary. The concept of record-linkage process first needs to be reviewed. Consider two files; file A and file B, both containing records, each record representing an individual. A record-linkage process brings together one record from file A with one record from file B. The records are compared, producing the comparison pattern γ . For the purpose of this paper, this comparison pattern is a vector $\gamma = [\gamma^1, \dots, \gamma^N]$, where N is the dimensionality of the vector. Each dimension corresponds to a comparison field recorded for each individual, such as last name, age, address, *etc.* With no loss of generality, γ^i is assigned the value 0 if the records disagree over comparison field i and it is assigned 1 if they agree. The comparison space Γ is assumed to be the set of all binary vectors (*i.e.* whose components are 0 or 1) of dimension N .

2.2 Underlying Probabilistic Processes

A record-linkage process is governed by an underlying probabilistic process. A good knowledge of the probabilistic process is needed to extract information from the record-linkage process. The formulation of the underlying probabilistic process is presented here in general terms. It is made more specific in the next section.

Consider a particular comparison pattern γ , define $m(\gamma)$ as the probability of observing γ , given that the two records producing γ , when brought together, represent the same individual. Similarly, define $u(\gamma)$ as the probability of observing γ , given that the two records producing γ , when brought together, do not represent the same individual. These two conditional probabilities, along with the probability of a match, define the underlying probabilistic process. The probabilistic process drives the record-linkage process. $m(\gamma)$ and $u(\gamma)$ are fundamental in the construction of record linkage rules; in particular most discriminatory rules. Record-linkage rules are devices to retrieve matches. They are defined next.

2.3 Record-Linkage Rules

In practice, a record-linkage rule classifies the pairs generated by a record-linkage process in one of three possible categories: a link, a non-link or a possible link. A link is an inferred matched pair and a non-link is an inferred unmatched pair. The pairs classified as possible links are set aside for further examination and eventually they are reclassified as links or non-links. The rule is based only on the value of the comparison vectors corresponding to each pair. The errors induced by a record-linkage rule are of two types: the type I error measuring the proportion

of unmatched pairs among the pairs classified as links under the linkage rule, and the type II error measuring the proportion of matches among the pairs classified as non-links.

The objective of record-linkage, from the standpoint of the paper, is to construct a most discriminatory record-linkage rule; that is one that will retrieve a maximum number of links while keeping the type I error under control. To accomplish this, let the comparison patterns be indexed according to decreasing value of $m(\gamma)/u(\gamma)$ to obtain the sequence $\{\gamma_1, \gamma_2, \dots, \gamma_m\}$, where M is the total number of pairs. Fellegi and Sunter (1969) show that the rule declaring the pairs whose index is smaller than some upper bound K “links” is the most discriminatory record linkage rule. The upper bound K is a function of the maximum type I error tolerated. The rule is most discriminatory in the sense that for the same tolerance on the type I error, it is impossible to find another rule which, in the long run, will retrieve more matched pairs. This fact is a direct application of the Neyman-Pearson Lemma (DeGroot 1986, pp. 444-445). Two uses of the Fellegi-Sunter rule are illustrated in section 6.

The Fellegi-Sunter record-linkage rule is articulated around the ratio $m(\gamma)/u(\gamma)$. Usually this ratio is estimated from the data through a model of the underlying probabilistic process. It is assumed that the model is a genuine representation of the probabilistic process. If the representation is not genuine, then substituting $m(\gamma)/u(\gamma)$ in the Fellegi-Sunter rule may not yield a most discriminatory record-linkage rule. Therefore, particular care must be taken in the choice of the model. The next section introduces models designed to describe the underlying probabilistic process in given situations.

3. MODELS FOR RECORD-LINKAGE

Two models formulating underlying probabilistic processes are presented in this section. The first model is a general formulation of any underlying process. The second model is an application of the first. In some situations, the second model is a good representation of the underlying probabilistic process and the Fellegi-Sunter rule based on this model is most discriminatory. Parameter estimation is discussed so that the expressions involved in the Fellegi-Sunter rule can be evaluated.

3.1 Latent Class Models

Because of the particular nature of a record-linkage process, the underlying probabilistic process can always be represented by a latent class model. A latent class model is built around latent variables. Generally speaking, a latent variable is a variable not observable, characterizing any observation generated by the probabilistic process. Latent variables classify the observations into latent

classes. In this problem, the observations are the comparison vectors (*i.e.* comparison patterns). An obvious latent variable categorizing the observations into two latent classes is the status of the pair associated with each comparison vector. This status is that of a matched pair status or of an unmatched pair status. The corresponding latent classes are the class of matched pairs and the class of unmatched pairs. A mathematical representation is given next to enable development of specific latent class models.

Let ν_{k,i_1,\dots,i_N} represent the count of pairs with the following attributes: if $k = 0$ the corresponding pairs have an unmatched pair status and if $k = 1$ they have a matched pair status. Furthermore, whenever $i_s = 0$, the corresponding pairs do not exhibit record agreement over the comparison field s and whenever $i_s = 1$, the pairs do exhibit record agreement over the comparison field s . Note that $s = 1, \dots, N$, where N is the number of comparison fields. It is important to keep in mind that the counts ν_{k,i_1,\dots,i_N} cannot be observed. Rather, what is observed are the counts aggregated over the latent classes. The aggregated counts are denoted by ν_{i_1,\dots,i_N} where

$$\nu_{i_1,\dots,i_N} = \nu_{0,i_1,\dots,i_N} + \nu_{1,i_1,\dots,i_N}. \quad (1)$$

While only the aggregated counts are observable in record-linkage situations, models are usually expressed in terms of the basic counts. This is done only for convenience. The following subsection is more specific and a simple latent class model for record linkage is introduced.

3.2 Conditional Independence

The conditional independence models are the simplest latent class models. Despite their simplicity, these models are an accurate representation of the underlying probabilistic process in some situations. Goodman (1974) gives a thorough analysis of several conditional independence models. Haberman (1979) gives a presentation of several conditional independence models, along with appropriate techniques of parameter estimation.

In this section, the conditional independence model for record linkage is introduced and its implications in terms of the underlying probabilistic process are exposed. The model is best described in its log-linear representation:

$$\log(\nu_{k,i_1,\dots,i_N}) = \mu + \lambda_k + \sum_{j=1}^N \alpha_{i_j}^j + \sum_{j=1}^N \zeta_{k,i_j}^j. \quad (2)$$

Naturally, there are constraints attached to the parameters of the model given in (2):

$$\begin{aligned} \lambda_1 &= -\lambda_0; \alpha_1^j = -\alpha_0^j; \zeta_{k,1}^j = -\zeta_{k,0}^j; \zeta_{1,i_j}^j = -\zeta_{0,i_j}^j, \\ k &= 0,1; j = 1, \dots, N; i_j = 0,1. \end{aligned} \quad (3)$$

The expression on the right-hand side of (2) includes one term for the latent variable (λ_k) and one term for each comparison field ($\alpha_{i_j}^j$). It also includes interaction terms (ζ_{k,i_j}^j). Each interaction is between a field and the latent variable. There are no direct interaction between the comparison fields. In other words, conditional on each latent class, agreements and disagreements over the comparison fields occur independently.

The assumption that the comparison variables are independent given the value of the latent variable is implicit when deriving inference through a conditional independence model. In practice, however, the underlying probabilistic process often conflicts with this assumption. Then the Fellegi-Sunter record-linkage rule constructed assuming model (2) may not be most discriminatory. In that situation, the discriminatory power can be raised through a better elicitation of the model. In fact, more elaborate latent class models integrate a higher degree of complexity in the relationships between the comparison fields themselves and between the comparison fields and the latent variable. These models can take a large number of forms according to the nature of a particular record-linkage situation. An instance of such models is presented in Section 4.

3.3 Parameter Estimation for the Conditional Independence Model

Once a model has been formulated, the values of its parameters must be evaluated. Then the Fellegi-Sunter rule is constructed from the model using the corresponding estimated values for $m(\gamma)$ and $u(\gamma)$. The parameter estimation process shall be reliable enough to prevent a significant loss of discriminatory power by way of the estimation error.

One feature of the latent class models makes them prone to estimation error: unidentifiability. Latent class models typically are unidentifiable in the sense that the equations maximizing the likelihood admit more than one solution. Parameter estimation with unidentifiable models remains difficult and confusing. However, from experience, the author found that for the conditional independence models, unidentifiability is usually not a determinant factor in the estimation error. A larger part of the error typically comes from the inadequacy of the model as a genuine representation of the underlying probabilistic process.

A suitable parameter-estimation technique for conditional independence models stems from approaching the problem as one of finding a maximum likelihood estimator in the presence of "missing observations". The missing observation in this case is the latent variable, the status of each pair. In the general context of parameter estimation with missing observations, Expectation-Maximization (E.M.) algorithms are quite popular. In fact, the E.M. algorithm is implemented without difficulty in the estimation

of the parameters of the conditional independence model given in (2) (Winkler 1988). But if there is considerable departure from the independence assumption, the value of the estimates becomes difficult to interpret (An example of this is given in section 4).

4. THE ST. LOUIS DATA: AN EXAMPLE OF A COMPLEX RECORD-LINKAGE PROCESS

This section introduces a particular example of a record-linkage process. A model is developed specifically to represent the underlying probabilistic process supporting this record-linkage process. It is expected that this model will induce more discrimination power in the application of the Fellegi-Sunter rule than the conditional independence model would.

4.1 Observable Latent Variable

The example is based on data collected in 1988 during a dress rehearsal in preparation for the Decennial Census Operations. Basically, there are two separate and presumably exhaustive surveys of all the individuals living in a defined geographical area within the city of St. Louis, Missouri. For each survey and for each individual available at the time of the survey, a record is created and various characteristics of the individual are recorded. These characteristics are: house number, phone number, street name, first name, last name, middle initial, marital status, age, race, sex, relationship with the respondent. The records of the two surveys are linked together.

For this particular application, the latent variable is made observable through an extensive follow-up study for the purpose of this and other researches. In the present situation, the information extracted from the latent variable leads to the construction of a model representative of the probabilistic process underlying the record-linkage process. Ultimately the discrimination power of this model is compared with that of the conditional independence model. The motivations leading to the construction of the model are presented in the following subsections.

4.2 Blocking and Dependencies

The goal of record-linkage is to retrieve as many matched pairs as possible given an upper bound on the type I error. The first obstacle is often the size of the files. The files may be quite large, making it impossible to examine all the pairs consisting of one record of file A and one record of file B. Blocking is considered whenever an exhaustive review of all the pairs is too costly and/or too time consuming.

The principle of blocking is as follows: To bring down the number of comparisons and other associated operations, the records of each file are assigned to blocks according to the value of a few key characteristics. These

characteristics are called the blocking variables. Only the records whose blocking variables take the same values may be brought into pairs. Since the records forming a matched pair tend to agree on the blocking characteristics, it is natural to expect the vast majority of the pairs discarded to be unmatched, as a result of the blocking scheme.

In the St. Louis example, the census file has 15,048 records, while the PES file contains 12,072 records. Potentially, there are over 180,000,000 pairs available for review. This number is excessive and blocking must be used to keep the size of the problem manageable. Therefore, the records are blocked on the first character of the surname and on a geographical unit called geocode. The geographical area encompassed by a given value of the geocode may consist of several street blocks, or two or more nearby perpendicular or parallel streets. This scheme yields blocks of reasonable sizes. Under this design, 116,305 pairs provide the information to construct inference.

Unfortunately, while it brings down the size of the problem, blocking on geocode also has undesirable side effects: it induces strong dependencies between the household variables among the unmatched pairs. The household variables are the last name, house number, street name and telephone number. For instance, consider two individuals forming an unmatched pair but who are part of the same block. Now, suppose these two individuals agree on the last name. Intuitively, given this information, chances are higher that the two individuals are from the same household. Therefore, the probabilities of agreement over the other household fields, given the information of agreement on the last name, are higher than the marginal probabilities. The nature of the dependencies between the household variables is studied next.

4.3 Measuring the Dependencies

To construct a model representative of the St. Louis record-linkage process, the dependencies between the household variables must be assessed. The information on the latent variable allows this. Table 1 gives the correlations of the responses of record comparisons over the comparison fields for the matched pairs. Table 2 gives the correlations of the responses of the record comparisons over the comparison fields for the unmatched pairs. For both matrices, all the correlations greater or equal to .01 are given. A correlation is not shown only if it is smaller than .01.

The correlations in Table 1, are rather small and overall do not suggest a significant pattern of dependency among the comparison variables restricted to the matched pairs. Note in particular that the correlations between the household variables are small among the matched pairs, suggesting little or no dependency. This can be explained by the fact that among the matched pairs, the agreement rate over any household field is very high and has a behavior close to that of a constant.

Table 1
Correlations Between Selected Comparison Fields
over the Set of Links

	Middle In.	Street	Phone	Marital
First Name	.123	0.	.045	.032
Middle In.	1	.010	.161	.079
House No.	.017	.194	.037	0.
Street	.01	1	.035	0.
Phone	.161	.035	1	.107
Age	.051	.004	.075	.118
Marital	.079	0.	.107	1

Table 2
Correlations Between Selected Comparison Fields
over the Set of Non-Links

	House No.	Street	Phone	Marital	Race
Last N.	.748	.326	.642	.099	.101
House No.	1	.400	.699	.111	.105
Street	.400	1	.292	.043	.086
Age	.104	.054	.086	.165	.024
Rel	.121	.068	.084	.394	.049

But in Table 2, the effects of blocking are evident in the high values of the correlations associated with the household variables restricted to the unmatched pairs. A sensible design for the model of the underlying probabilistic process should account for these high correlations by incorporating dependency components.

4.4 A Model Tailored for the St. Louis Data

In order to make valid inference on the status of the pairs, a model descriptive of the underlying probabilistic process must be elicited. The conditional independence model presented in (2) is attractive because of its simplicity. However, it is clear at this point that this model does not correctly represent the probabilistic process underlying the St. Louis record-linkage process. An educated model is introduced, motivated by the information made available on the dependencies between the household variables.

To appreciate the more general structure of the educated model, some conventions must be set regarding the indexing of the comparison fields: comparison field 1 is the last name, comparison field 2 is the house number, comparison field 3 is the street name, and comparison field 4 is the phone number. The seven remaining comparison fields are indexed arbitrarily by the values 5-11. The educated model accounts for all possible interaction effects between fields 1 through 4 among the unmatched pairs. The log-linear representation of the educated model is as follows:

$$\log(\nu_{k,i_1,\dots,i_{11}}) = \mu + \lambda_k + \sum_{j=1}^{11} \alpha_{ij}^j + \sum_{j=1}^{11} \zeta_{k,ij}^j + (1-k) \left(\sum_{\{j < l \leq 4\}} \eta_{ij,il}^{j,l} + \sum_{\{1 \leq j < l < m \leq 4\}} \Phi_{ij,il,im}^{j,l,m} + \Psi_{i_1,i_2,i_3,i_4}^{1,2,3,4} \right). \quad (4)$$

Note the coefficient $(1 - k)$ multiplying the household interaction terms, indicating that the dependency relation between the household variables is only among the unmatched pairs. This contrasts with the symmetry of the conditional independence model in (2).

The restrictions in (3) apply here as well. In addition, more constraints must be satisfied. The following constraints are imposed on the interaction terms of the second order:

$$\eta_{ij,1}^{j,l} = -\eta_{ij,0}^{j,l}; \quad \eta_{1,il}^{j,l} = -\eta_{0,il}^{j,l}. \quad (5)$$

The range of the indices is $1 \leq j < l \leq 4$. The constraints on the interaction terms of the third order are:

$$\begin{aligned} \Phi_{ij,il,1}^{j,l,m} &= -\Phi_{ij,il,0}^{j,l,m}; & \Phi_{ij,1,im}^{j,l,m} &= -\Phi_{ij,0,im}^{j,l,m}; \\ \Phi_{1,il,im}^{j,l,m} &= -\Phi_{0,il,im}^{j,l,m}. \end{aligned} \quad (6)$$

The range of the indices in this case is: $1 \leq j < l < m \leq 4$. Finally, the constraints on the fourth order interaction terms are:

$$\begin{aligned} \Psi_{i_1,i_2,i_3,1}^{1,2,3,4} &= -\Psi_{i_1,i_2,i_3,0}^{1,2,3,4}; & \Psi_{i_1,i_2,1,i_4}^{1,2,3,4} &= -\Psi_{i_1,i_2,0,i_4}^{1,2,3,4}; \\ \Psi_{i_1,1,i_3,i_4}^{1,2,3,4} &= -\Psi_{i_1,0,i_3,i_4}^{1,2,3,4}; & \Psi_{1,i_2,i_3,i_4}^{1,2,3,4} &= -\Psi_{0,i_2,i_3,i_4}^{1,2,3,4}. \end{aligned} \quad (7)$$

It is natural to expect the educated model (4) to be more discriminatory since it accounts for interactions between the household variables. In section 6, the performances of the two models are presented.

4.5 Parameter Estimation for Models with Dependencies

Parameter estimation for models with dependencies is far more difficult than for conditional independence models. For the St. Louis example, the scoring algorithm given by Haberman (1979, p. 547) was used to estimate the parameters of the educated model (4). This technique can be regarded as an E.M. algorithm where the maximization part (M. step) is an application of the Newton-Raphson algorithm.

The most important difficulty when using this technique is the choice of a starting point. The following strategy is adopted to choose a starting point. First, the parameters of the conditional independence model (2) are estimated via the E.M. algorithm presented in subsection 3.3. Then an intermediate model is constructed. The intermediate model, in this case, embeds all the second and lower order interaction terms of the educated model (4). The estimated parameters of the conditional independence model can serve to construct the starting point to estimate the parameters of the intermediate model through the scoring algorithm. Finally, the estimates of the parameters of the intermediate model are used as a starting point to estimate the parameters of the educated model (4), via the scoring algorithm.

5. THE AD-HOC APPROACH

In the last section, a complex model representing an underlying probabilistic process was elicited for the St. Louis data. In this situation, the elicitation is easy since follow-up information is available. Of course in practice, follow-up information is not available. It is often too difficult and/or too expensive to go through the elicitation and estimation procedures to determine the structure of the underlying process and the values of the parameters. In those cases, an ad-hoc approach might be appropriate. In the St. Louis example, the ad-hoc approach consists of adjusting the parameters of the process derived from the conditional independence model (2) to obtain a more discriminatory model.

Note that under both model (2) and model (4), for the matched pair, the agreement or disagreements over the comparison fields are independent. This means that the following formula applies in both situations.

$$m(\gamma) = \prod_{i=1}^N m_i^{x_i} (1 - m_i)^{1-x_i},$$

m_i is the probability of agreement over field i of two records forming a matched pair. Furthermore, $x_i = 0$ if the pattern γ calls for a disagreement over field i and $x_i = 1$ if it calls for an agreement. The idea behind the ad-hoc method is to keep the conditional independence structure in (2), but to adjust the values of the m_i 's.

The probabilities of agreement, conditional on a matched pair, evaluated under the conditional independence model and the educated model are given in Table 3. The difference between the probability corresponding to the educated model with the probability corresponding to the conditional independence model can be quite substantial for some fields. In particular, the difference is important in the case of the first name field.

Table 3
Probabilities of Agreement Conditional
on a Matched Pair

Comparison Field	Cond'l Indep.	Educated
Last Name	.9430	.9561
First Name	.3319	.9140
Mid. Init.	.2125	.5222
House No.	.9692	.9724
Street Name	.9179	.9194
Phone	.6619	.6887
Age	.3903	.8602
Relation	.3353	.4986
Marital Status	.6072	.8547
Sex	.6134	.4842
Race	.9672	.9018

In general, experience shows that the conditional probability of agreement over first name, conditional on a matched pair, is around .99, closer to the .91 value obtained under the educated model. Therefore, after estimating the parameters of the conditional independence model through the E.M. algorithm, the probability of agreement over the first name given a match status is replaced by the value .99. The probability of agreement over the last name given a matched pair is also replaced by the value .99. This procedure increases the discriminatory power associated with the conditional independence model in the application of the Fellegi-Sunter rule.

6. APPLYING THE FELLEGI-SUNTER RULE

6.1 St.Louis

This subsection evaluates the discrimination power of the Fellegi-Sunter rule when applied to the St-Louis record-linkage data and assuming, in turn, three different underlying probabilistic processes. The three underlying probabilistic processes assumed are derived directly from the conditional model (2), directly from the educated model (4), and finally, from the conditional model (2), through the ad-hoc procedure. The following table gives a comparative measure of the performance of the Fellegi-Sunter rule under each of the 3 assumptions regarding the underlying process. The performance is evaluated making use of the privileged information available on the latent variable.

Each cell of Table 4 contains three entries. The first of these entries is the number of matched pairs that were designated links through the Fellegi-Sunter record-linkage rule, assuming each of the three underlying processes, and under four different controlled Type I errors. The total number of matched pairs that could theoretically be recovered is 9,823. The second entry of each cell is the total number of pairs designated link through the Fellegi-Sunter rule. The third entry of the cell is the upper bound on the

Type I error. Recall that the Fellegi-Sunter rule maximizes the number of links under a fixed type I error provided it is based on the correct underlying process. The first column of Table 4 gives the counts assuming an underlying process derived from the conditional independence model (2). The second column gives the same quantities assuming an underlying process derived from the educated model (4). Finally, the third column gives the same numbers assuming an underlying process derived from the conditional independence model and adjusted through the ad-hoc procedure.

Table 4

St. Louis: Links Recovered via Three Approaches
under Four Error Levels

	Independence Assumption	Household Interactions	Ad-hoc Procedure
Links	6,404	9,012	6,476
Pairs	6,436	9,056	6,508
Error Bound	.005	.005	.005
Links	7,273	9,712	9,562
Pairs	7,346	9,808	9,659
Error Bound	.01	.01	.01
Links	9,636	9,758	9,765
Pairs	9,824	9,952	9,960
Error Bound	.02	.02	.02
Links	9,740	9,776	9,783
Pairs	10,038	10,062	10,097
Error Bound	.03	.03	.03

There are two important facts that can be deduced from this table. First, the rule based on an underlying process derived from the educated model (4) does consistently better than the rule based on an underlying process derived from the conditional independence model in terms of matches retrieved. Secondly, the performances of the rules differ most when the bound on the type I error is small and at that level (.005), the rule based on an underlying probability process derived from the educated model is clearly superior. When the bound is larger (.03), the underlying probabilistic models are more or less equivalent in terms of induced discrimination power.

6.2 Columbia

The same type of data were collected throughout the area of Columbia, Missouri. The data are slightly different because some of the records have a rural format, that is the street name is replaced by the rural route number and the house number by the box number. Nevertheless, the same relations of dependencies emerge and the same model is appropriate. Table 5 gives a summary of the discrimination achieved at 2 levels of tolerance on the type I error. Taking into account the blocking scheme, there are 6,780 retrievable pairs.

Table 5

Columbia: Links Recovered via Three Approaches
under Two Error Levels

	Independence Assumption	Household Interactions	Ad-hoc Procedure
Links	700	1,268	2,035
Pairs	704	1,276	2,046
Type I Error	.005	.005	.005
Links	5,954	6,607	6,545
Pairs	6,016	6,675	6,612
Type I Error	.01	.01	.01

In the case of Columbia, it is clear again than the educated model does better than the conditional independence model. It should be noted that in practice, the ad-hoc approach built on the conditional independence model performs as well as the educated model. The educated model however, is preferred because of its sound theoretical basis.

7. A SUGGESTION FROM AN ANONYMOUS REFEREE

Another ad-hoc technique is suggested by an anonymous referee. The referee points out that a large majority of the pairs examined in situations like these are unmatched. In the case of St. Louis, 91.5% of the pairs examined turn out to be unmatched. Given this proportion, the trends animating the comparison variables over the set of all pairs, mostly reflect the activity of the unmatched pairs. This reasoning can be extended further to conclude that the estimation of the parameters of the dependency structure underlying the unmatched pairs can be carried through successfully by treating the set of all pairs as if it were the set of unmatched pairs. The parameter estimation becomes trivial. The parameters that must be estimated characterize a simple log-linear model, without any latent variable (Fienberg, Bishop and Holland, p. 24). The parameters descriptive of the matches can be estimated separately through a simple iterative technique such as the E.M. algorithm, combined with *a priori* information.

The approach of the referee does proceed from a realistic model of the process, and in that way, it is in agreement with the thrust of this paper. But the effort of the paper is also to devise discriminatory rules, while sticking to the latent structure constraint. In situations where the proportion of matched pairs is high, or dependencies are manifest among the matches, the approach of the referee fails. A parameter estimation derived directly from the natural model, if feasible, is recommended.

8. CONCLUSIONS

The goal of the research was to show how a better elicitation of the probabilistic models supporting record-linkage processes can induce accrued discriminatory power in the Fellegi-Sunter record-linkage rule. In the cases of the St. Louis and Columbia examples, this goal was certainly achieved. The educated model given in (4) is indeed more descriptive of the underlying probabilistic process and it induces a good deal more discrimination power in the Fellegi-Sunter rule than the conditional independence model (2).

The techniques used for the St. Louis and Columbia data can also be used for the analysis of other data set generated by record-linkage processes supported by a probabilistic process with a similar dependency structure. This dependency structure is certain to surface in any record-linkage application involving the matching of records of individuals on a set of household variables (last name, street name, house number, phone, rural address *etc.*). It is also likely to occur when matching records of businesses on household variables.

There are two major difficulties in the way, when seeking improved discriminatory power by model elicitation. First, since the probability structure underlying the process is usually unknown, to elicit the structure or the corresponding statistical model involves a considerable investigative effort and the cost involved may be prohibitive. Second, even assuming that the correct model is available, the estimation procedures available for the parameter estimation are difficult to handle and poorly understood. More research and work are needed to understand and, to a degree, overcome these two difficulties.

It must also be pointed out that methods based on ad-hoc adjustments of the type described in section 5, and on approximations, as suggested by an anonymous referee, also increase the discriminatory power of the Fellegi-Sunter rule substantially in situations of the type of St. Louis or Columbia. Techniques of this type are serious competitors. The parameter estimation is easy and the associated Fellegi-Sunter rule can be just as crisp in some cases. However, the assumptions supporting these techniques are flawed and the resulting Fellegi-Sunter rule is pathological, providing an unsteady basis on which to make decisions. A model with parameters estimated "naturally" is preferable. The ad-hoc techniques and approximations are recommended when the elicitation of an educated model seems not possible, or the estimation of the parameters of the educated model appears excessively difficult.

A word must be said about the St. Louis and Columbia data. These data are of very high quality. This explains in part the very successful rate of matching exhibited in both the St. Louis and Columbia examples. It is also reasonable to expect a less clear-cut difference between the various linkage techniques had the data been lower quality.

ACKNOWLEDGEMENTS

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author is grateful to the anonymous referee for his/her patience and constructive suggestions. The author is also indebted to William E. Winkler for the guidance he provided throughout the learning process that led to this paper.

REFERENCES

- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- DeGROOT, M.H. (1986). *Probability and Statistics*, 2nd. Edition. Reading, MA: Addison-Wesley.
- FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 40, 1183-1210.
- GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 2, 215-231.
- HABERMAN, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press.
- HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.
- THIBAudeau, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.
- WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Bureau of The Census Fifth Annual Research Conference*, 145-155.
- WINKLER, W.E. (1988). Using The E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

Regression Analysis of Data Files that are Computer Matched

FRITZ SCHEUREN and WILLIAM E. WINKLER¹

ABSTRACT

This paper focuses on how to deal with record linkage errors when engaged in regression analysis. Recent work by Rubin and Belin (1991) and by Winkler and Thibaudeau (1991) provides the theory, computational algorithms, and software necessary for estimating matching probabilities. These advances allow us to update the work of Neter, Maynes, and Ramanathan (1965). Adjustment procedures are outlined and some successful simulations are described. Our results are preliminary and intended largely to stimulate further work.

KEY WORDS: Record linkage; Matching error; Regression analysis.

1. INTRODUCTION

Information that resides in two separate computer data bases can be combined for analysis and policy decisions. For instance, an epidemiologist might wish to evaluate the effect of a new cancer treatment by matching information from a collection of medical case studies against a death registry in order to obtain information about the cause and date of death (*e.g.*, Beebe 1985). An economist might wish to evaluate energy policy decisions by matching a data base containing fuel and commodity information for a set of companies against a data base containing the values and types of goods produced by the companies (*e.g.*, Winkler 1985). If unique identifiers, such as verified social security numbers or employer identification numbers, are available, then matching data sources can be straightforward and standard methods of statistical analysis may be applicable directly.

When unique identifiers are not available (*e.g.*, Jabine and Scheuren 1986), then the linkage must be performed using information such as company or individual name, address, age, and other descriptive items. Even when typographical variations and errors are absent, name information such as “Smith” and “Robert” may not be sufficient, by itself, to identify an individual. Furthermore, the use of addresses is often subject to formatting errors because existing parsing or standardization software does not effectively allow comparison of, say, a house number with a house number and a street name with a street name. The addresses of an individual we wish to match may also differ because one is erroneous or because the individual has moved.

Over the last few years, there has been an outpouring of new work on record linkage techniques in North America (*e.g.*, Jaro 1989; and Newcombe, Fair and Lalonde 1992). Some of these results were spurred on by

a series of conferences beginning in the mid-1980s (*e.g.*, Kilss and Alvey 1985; Howe and Spasoff 1986; Coombs and Singh 1987; Carpenter and Fair 1989); a further major stimulus in the U.S. has been the effort to study under-coverage in the 1990 Decennial Census (*e.g.*, Winkler and Thibaudeau 1991). The new book by Newcombe (1988) has also had an important role in this ferment. Finally, efforts elsewhere have also been considerable (*e.g.*, Copas and Hilton 1990).

What is surprising about all of this recent work is that the main theoretical underpinnings for computer-oriented matching methods are quite mature. Sound practice dates back at least to the 1950s and the work of Newcombe and his collaborators (*e.g.*, Newcombe *et al.* 1959). About a decade later, the underlying theory for these basic ideas was firmly established with the papers of Tepping (1968) and, especially, Fellegi and Sunter (1969).

Part of the reason for the continuing interest in record linkage is that the computer revolution has made possible better and better techniques. The proliferation of machine readable files has also widened the range of application. Still another factor has been the need to build bridges between the relatively narrow (even obscure) field of computer matching and the rest of statistics (*e.g.*, Scheuren 1985). Our present paper falls under this last category and is intended to look at what is special about regression analyses with matched data sets.

By and large we will not discuss linkage techniques here. Instead, we will discuss what happens *after* the link status has been determined. The setting, we will assume, is the typical one where the linker does his or her work separately from the analyst. We will also suppose that the analyst (or user) may want to apply conventional statistical techniques – regression, contingency tables, life tables, *etc.* – to the linked file. A key question we want to explore then is “What should the linker do to help the analyst?” A

¹ Fritz Scheuren, U.S. Internal Revenue Service, Washington DC 20224; William E. Winkler, U.S. Bureau of the Census, Washington DC 20233.

related question is “What should the analyst know about the linkage and how should that information be used?”

In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analyses done. Despite this, rarely are we given direct measures of that impact (e.g., Scheuren and Oh 1975). Rubin (1990) has noted the need to make inferential statements that are designed to summarize evidence in the data being analyzed. Rubin’s ideas were presented in the connotation of data housekeeping techniques like editing and imputation, where nonresponse can often invalidate standard statistical procedures that are available in existing software packages. We believe Rubin’s perspective applies at least with equal force in record linkage work.

Organizationally, our discussion is divided into four sections. First, we provide some background on the linkage setting, because any answers – even partial ones – will depend on the files to be linked and the uses of the matched data. In the next section we discuss our methodological approach, focusing, as already noted, just on regression analysis. A few results are presented in section 4 from some exploratory simulations. These simulations are intended to help the reader weigh our ideas and get a feel for some of the difficulties. A final section consists of preliminary conclusions and ideas for future research. A short appendix containing more on theoretical considerations is also provided.

2. RECORD LINKAGE BACKGROUND

When linking two or more files, an individual record on one file may not be linked with the correct corresponding record on the other file. If a unique identifier for corresponding records on two files is not available – or is subject to inaccuracy – then the matching process is subject to error. If the resultant linked data base contains a substantial proportion of information from pairs of records that have been brought together erroneously or a significant proportion of records that need to be brought together are erroneously left apart, then statistical analyses may be sufficiently compromised that results of standard statistical techniques could be misleading. For the bulk of this paper we will only be treating the situation of how erroneous links affect analyses. The impact of problems caused by erroneous nonlinks (an implicit type of sampling that can yield selection biases) is discussed briefly in the final section.

2.1 Fellegi-Sunter Record Linkage Model

The record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (e.g.,

Newcombe *et al.* 1959), Fellegi and Sunter (1969) considered ratios of probabilities of the form:

$$R = Pr(\gamma \in \Gamma \mid M) / Pr(\gamma \in \Gamma \mid U), \quad (2.1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ . For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Smith or Zabrinsky, occur. The fields that are compared (surname, first name, age) are referred to as *matching variables*.

The decision rule is given by:

If $R > Upper$, then designate pair as a link.

If $Lower \leq R \leq Upper$, then designate pair as a possible link and hold for clerical review. (2.2)

If $R < Lower$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that the decision rule is optimal in the sense that for any pair of fixed bounds on R , the middle region is minimized over all decision rules on the same comparison space Γ . The cutoff thresholds *Upper* and *Lower* are determined by the error bounds. We call the ratio R or any monotonely increasing transformation of it (such as given by a logarithm) a *matching weight* or *total agreement weight*.

In actual applications, the optimality of the decision rule (2.2) is heavily dependent on the accuracy of the estimates of the probabilities given in (2.1). The probabilities in (2.1) are called *matching parameters*. Estimated parameters are (nearly) *optimal* if they yield decision rules that perform (nearly) as well as rule (2.2) does when the true parameters are used.

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to record linkage. To describe the model further, suppose there are two files of size n and m where – without loss of generality – we will assume that $n \leq m$. As part of the linkage process, a comparison might be carried out between all possible $n \times m$ pairs of records (one component of the pair coming from each file). A decision is, then, made as to whether or not the members of each comparison-pair represent the same unit or whether there is insufficient evidence to determine link status.

Schematically, it is conventional to look at the $n \times m$ pairs arrayed by some measure of the probability that the pair represent records for the same unit. In Figure 1, for example, we have plotted two curves. The curve on the right is a hypothetical distribution of the n true links by the “matching weight” (computed from (2.1) but in natural logarithms). The curve on the left is the remaining of the $n \times (m - 1)$ pairs – the true nonlinks – plotted by their matching weights again in logarithms.

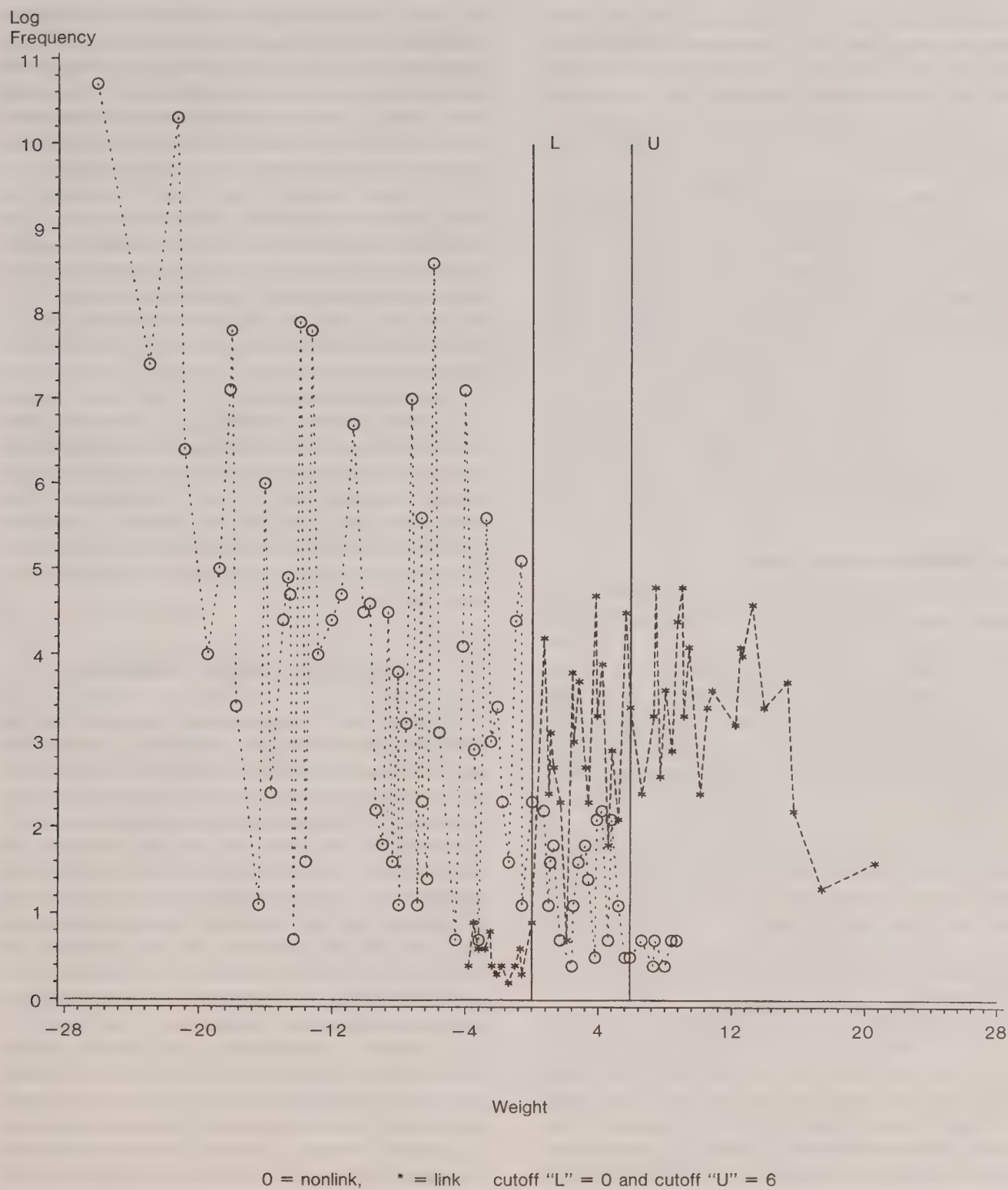


Figure 1. Log Frequency vs Weight, Links and Nonlinks

Typically, as Figure 1 indicates, the link and nonlink distributions overlap. At the extremes the overlap is of no consequence in arriving at linkage decisions; however, there is a middle region of potential links, say between “L” and “U”, where it would be hard, based on Figure 1 alone, to distinguish with any degree of accuracy between links and nonlinks.

The Fellegi-Sunter model is valid on any set of pairs we consider. However, for computational convenience, rather than consider all possible pairs in $\mathbf{A} \times \mathbf{B}$, we might consider only a subset of pairs where the records from both files agree on key or “blocking” information that is thought to be highly accurate. Examples of the *logical blocking criteria* include items such as a geographical identifier like Postal (e.g., ZIP) code or a surname identifier such as a Soundex or NYSIIS code (see e.g., Newcombe 1988, pp. 182-184). Incidentally, the Fellegi-Sunter Model does not presuppose (as Figure 1 did) that among the $n \times m$ pairs there will be n links but rather, if there are no duplicates on A or B, that there will be at most n links.

2.2 Handling Potential Links

Even when a computer matching system uses the Fellegi-Sunter decision rule to designate some pairs as almost certain *true links* or *true nonlinks*, it could leave a large subset of pairs that are only potential links. One way to address potentially linked pairs is to clerically review them in an attempt to delineate true links correctly. A way to deal with erroneously nonlinked pairs is to perform additional (again possibly clerical) searches. Both of these approaches are costly, time-consuming, and subject to error.

Not surprisingly, the main focus of record linkage research since the beginning work of Newcombe has been how to reduce the clerical review steps caused by the potential links. Great progress has been made in improving linkage rules through better utilization of information in pairs of records and at estimating error rates via probabilistic models.

Record linkage decision rules have been improved through a variety of methods. To deal with minor typographical errors such as “Smith” versus “Smoth”, Winkler and Thibaudeau (1991) extended the string comparator metrics introduced by Jaro (1989). Alternatively, Newcombe *et al.* (1989) developed methods for creating and using partial agreement tables. For certain classes of files, Winkler and Thibaudeau (1991) (see also Winkler 1992; Jaro 1989) developed Expectation-Maximization procedures and *ad hoc* modelling procedures based on *a priori* information that automatically yielded the optimal parameters in (2.1) for use in the decision rules (2.2).

Rubin and Belin (1991) introduced a method for estimating error rates, when error rates could not be reliably estimated via conventional methods (Belin 1991,

pp. 19-20). Using a model that specified that the curves of weights versus log frequency produced by the matching process could be expressed as a mixture of two curves (links and nonlinks), Rubin and Belin estimated the curves which, in turn, gave estimates of error rates. To apply their method, Rubin and Belin needed a training sample to yield an *a priori* estimate of the shape of the two curves.

While many linkage problems arise in retrospective, often epidemiological settings, occasionally linkers have been able to designate what information is needed in both data sets to be linked based on known analytic needs. Requiring better matching information, such as was done with the 1990 Census Post-Enumeration Survey (see e.g., Winkler and Thibaudeau 1991), assured that sets of potential links were minimized.

Despite these strides, eventually, the linker and analyst still may have to face a possible clerical review step. Even today, the remaining costs in time, money and hidden residual errors can still be considerable. Are there safe alternatives short of a full review? We believe so and this belief motivates our perspective in section 3, where we examine linkage errors in a regression analysis context. Other approaches, however, might be needed for different analytical frameworks.

3. REGRESSION WITH LINKED DATA

Our discussion of regression will presuppose that the linker has helped the analyst by providing a combined data file consisting of pairs of records – one from each input file – along with the match probability and the link status of each pair. Link, nonlink, and potential links would all be included and identified as such. Keeping likely links and potential links seems an obvious step; keeping likely nonlinks, less so. However, as Newcombe has pointed out, information from likely nonlinks is needed for computing biases. We conjecture that it will suffice to keep no more than two or three pairs of matches from the B file for each record on the A file. The two or three pairs with the highest matching weights would be retained.

In particular, we will assume that the file of linked cases has been augmented so that every record on the smaller of the two files has been paired with, say, the *two* records on the larger file having the highest matching weights. As $n \leq m$, we are keeping $2n$ of the $n \times m$ possible pairs. For each record we keep the linkage indicators and the probabilities associated with the records to which it is paired. Some of these cases will consist of (link, nonlink) combinations or (nonlink, nonlink) combinations. For simplicity's sake, we are not going to deal with settings where more than one true link could occur; hence, (link, link) combinations are by definition ruled out.

As may be quite apparent, such a data structure allows different methods of analysis. For example, we can partition

the file back into three parts – identified links, nonlinks, and potential links. Whatever analysis we are doing could be repeated separately for each group or for subsets of these groups. In the application here, we will use nonlinks to adjust the potential links, and, thereby, gain an additional perspective that could lead to reductions in the Mean Square Error (MSE) over statistics calculated only from the linked data.

For statistical analyses, if we were to use only data arising from pairs of records that we were highly confident were links, then we might be throwing away much additional information from the set of potentially linked pairs, which, as a subset, could contain as many true links as the set of pairs which we designate as links. Additionally, we could seriously bias results because certain subsets of the true links that we might be interested in might reside primarily in the set of potential links. For instance, if we were considering affirmative action and income questions, certain records (such as those associated with lower income individuals) might be more difficult to match using name and address information and, thus, might be heavily concentrated among the set of potential links.

3.1 Motivating Theory

Neter, Maynes, and Ramanathan (1965) recognized that errors introduced during the matching process could adversely affect analyses based on the resultant linked files. To show how the ideas of Neter *et al.* motivate the ideas in this paper, we provide additional details of their model. Neter *et al.* assumed that the set of records from one file (1) always could be matched, (2) always had the same probability p of being correctly matched, and (3) had the same probability q of being mismatched to any remaining records in the second file (i.e. $p + (N - 1)q = 1$ where N is file size). They generalized their basic results by assuming that the sets of pairs from the two files could be partitioned into classes in which (1), (2) and (3) held.

Our approach follows that of Neter *et al.* because we believe their approach is sensible. We concur with their results showing that if matching errors are moderate then regression coefficients could be severely biased. We do not believe, however, that condition (3) – which was their main means of simplifying computational formulas – will ever hold in practice. If matching is based on unique identifiers such as social security numbers subject to typographical error, it is unlikely that a typographical error will mean that a given record has the same probability of being incorrectly matched to all remaining records in the second file. If matching variables consist of name and address information (which is often subject to substantially greater typographical error), then condition (3) is even more unlikely to hold.

To fix ideas on how our work builds on and generalizes results of Neter *et al.* we consider a special case. Suppose

we are conducting ordinary least squares using a simple regression of the form,

$$y = a_0 + a_1x + \epsilon. \quad (3.1)$$

Next, assume mismatches have occurred, so that the y variables (from one file) and the x variables (from another file) are *not* always for the *same unit*.

Now in this setting, the unadjusted estimator of a_1 would be biased; however, under assumptions such as that x and y are independent when a mismatch occurs, it can be shown that, if we know the mismatch rate, h , that an unbiased adjusted estimator can be obtained by simply correcting the ordinary estimator by multiplying it by $(1/(1 - h))$. Intuitively, the erroneously linked pairs lead to an understatement of the true correlation (positive or negative) between x and y . The adjusted coefficient removes this understatement. With the adjusted slope coefficient \hat{a}_1 , the proper intercept can be obtained from the usual expression $\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}$, where \hat{a}_1 has been adjusted.

Methods for estimating regression standard errors can also be devised in the presence of matching errors. Rather than just continuing to discuss this special case, though, we will look at how the idea of making a multiplicative adjustment can be generalized. Consider

$$Y = X\beta + \epsilon, \quad (3.2)$$

the ordinary univariate regression model, for which error terms all have mean zero and are independent with constant variance σ^2 . If we were working with a data base of size n , Y would be regressed on X in the usual manner. Now, given that each case has two matches, we have $2n$ pairs altogether. We wish to use (X_i, Y_i) , but instead use (X_i, Z_i) . Z_i could be Y_i , but may take some other value, Y_j , due to matching error.

For $i = 1, \dots, n$,

$$Z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij} \text{ for } j \neq i, \end{cases} \quad (3.3)$$

$$p_i + \sum_j q_{ij} = 1.$$

The probability p_i may be zero or one. We define $h_i = 1 - p_i$ and divide the set of pairs into n mutually exclusive classes. The classes are determined by records from one of the files. Each class consists of the independent x -variable X_i , the true value of the dependent y -variable, the values of the y -variables from records in the second file to which the record in the first file containing X_i have been paired, and computer matching probabilities (or weights). Included are links, nonlinks, and potential links. Under an assumption of one-to-one matching, for each $i = 1, \dots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$.

The intuitive idea of our approach (and that of Neter *et al.*) is that we can, under the model assumptions, express each observed data point pair (X, Z) in terms of the true values (X, Y) and a bias term (X, b) . All equations needed for the usual regression techniques can then be obtained. Our computational formulas are much more complicated than those of Neter *et al.* because their strong assumption (3) made considerable simplification possible in the computational formulas. In particular, under their model assumptions, Neter *et al.* proved that both the mean and variance of the observed Z -values were necessarily equal the mean and variance of the true Y -values.

Under the model of this paper, we observe (see Appendix) that

$$\begin{aligned} E(Z) &= (1/n) \sum_i E(Z|i) = (1/n) \sum_i (Y_i p_i + \sum_j Y_j q_{ij}) \\ &= (1/n) \sum_i Y_i + (1/n) \sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i] \\ &= \bar{Y} + B. \end{aligned} \quad (3.4)$$

As each $X_i, i = 1, \dots, n$, can be paired with either Y_i or $Y_{\phi(i)}$, the second equality in (3.4) represents $2n$ points. Similarly, we can represent σ_{zy} in terms of σ_{xy} and a bias term B_{xy} , and σ_z^2 in terms of σ_y^2 and a bias term B_{yy} . We neither assume that the bias terms have expectation zero nor that they are uncorrelated with the observed data.

With the different representations, we can adjust the regression coefficients β_{zx} and their associated standard errors back to the true values β_{yx} and their associated standard errors. Our assumption of one-to-one matching (which is not needed for the general theory) is done for computational tractability and to reduce the number of records and amount of information that must be tracked during the matching process.

In implementing the adjustments, we make two crucial assumptions. The first is that, for $i = 1, \dots, n$, we can accurately estimate the true probabilities of a match p_i . See Appendix for the method of Rubin and Belin (1991). The second is that, for each $i = 1, \dots, n$, the true value Y_i associated with independent variable X_i is the pair with the highest matching weight and the false value $Y_{\phi(i)}$ is associated with the second highest matching weight. (From the simulations conducted it appears that at least the first of these two assumptions matters greatly when a significant portion of the pairs are potential links.)

3.2 Simulated Application

Using the methods just described, we attempted a simulation with real data. Our basic approach was to take two files for which true linkage statuses were known and re-link them using different matching variables – or really versions of the same variables with different degrees of distortion introduced, making it harder and harder to

distinguish a link from a nonlink. This created a setting where there was enough discrimination power for the Rubin-Belin algorithm for estimating probabilities to work, but not so much discriminating power that the overlap area of potential links becomes insignificant.

The basic simulation results were obtained by starting with a pair of files of size 10,000 that had good information for matching and for which true match status was known. To conduct the simulations a range of error was introduced into the matching variables, different amounts of data were used for matching, and greater deviations from optimal matching probabilities were allowed.

Three matching scenarios were considered: (1) *good*, (2) *mediocre*, and (3) *poor*. The good matching scenario consisted of using most of the available procedures that had been developed for matching during the 1990 U.S. Census (e.g., Winkler and Thibaudeau 1991). Matching variables consisted of last name, first name, middle initial, house number, street name, apartment or unit identifier, telephone, age, marital status, relationship to head of household, sex, and race. Matching probabilities used in crucial likelihood ratios needed for the decision rules were chosen close to optimal.

The mediocre matching scenario consisted of using last name, first name, middle initial, two address variations, apartment or unit identifier, and age. Minor typographical errors were introduced independently into one seventh of the last names and one fifth of the first names. Matching probabilities were chosen to deviate from optimal but were still considered to be consistent with those that might be selected by an experienced computer matching expert.

The poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. For the good scenario, we see that the scatter for true links and nonlinks is almost completely separated (Figure 2). With the mediocre scheme, the corresponding sets of points overlap moderately (Figure 3); and, with the poor, the overlap is substantial (Figure 4).

We primarily caused the good matching scenario to degenerate to the poor matching error (Figures 2-4) by using less matching information and inducing typographical error in the matching variables. Even if we had kept the same matching variables as in the good matching scenario (Figure 2), we could have caused curve overlap (as in Figure 4) merely by varying the matching

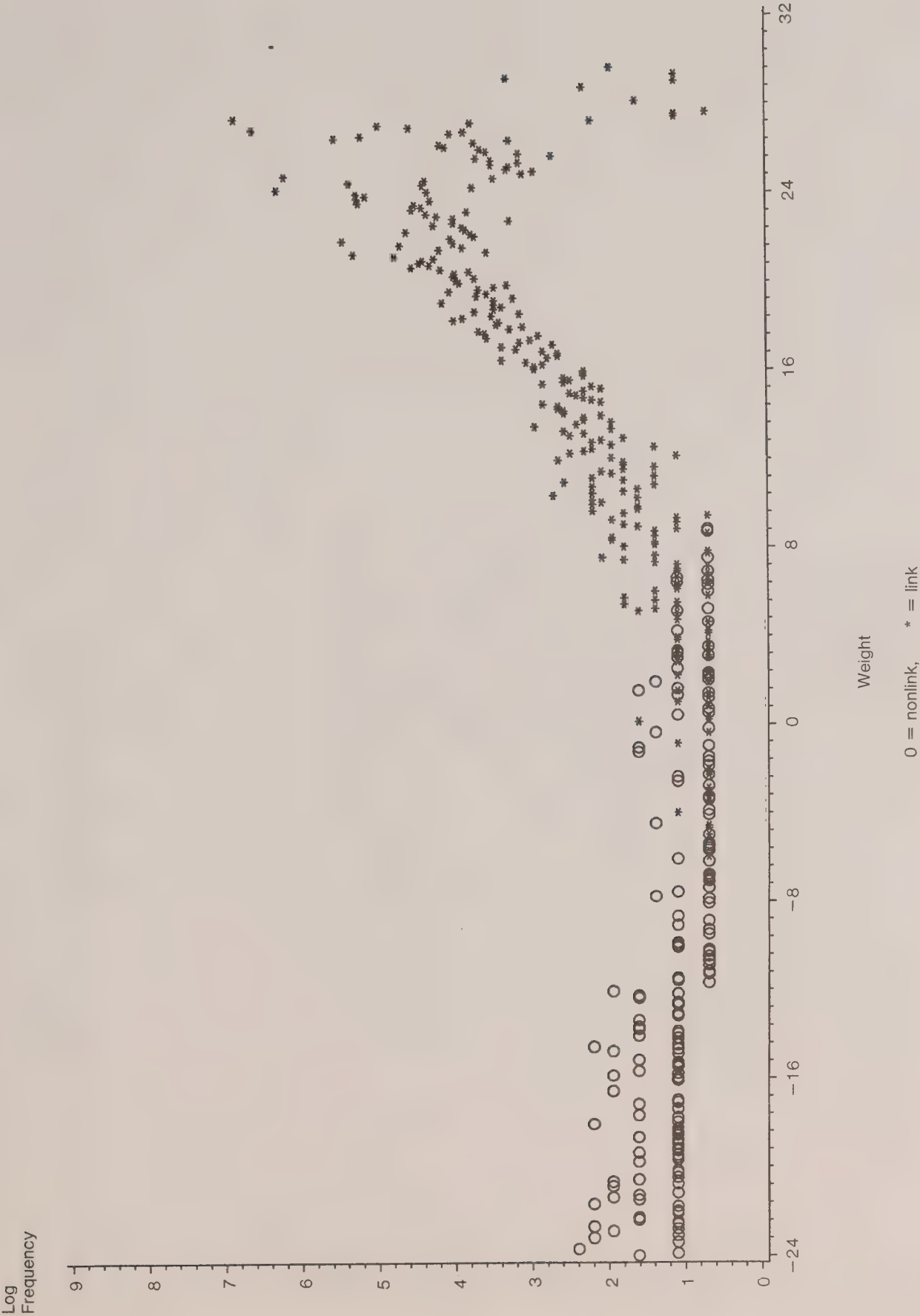


Figure 2. Log of Frequency vs Weight Good Matching Scenario, Links and Nonlinks

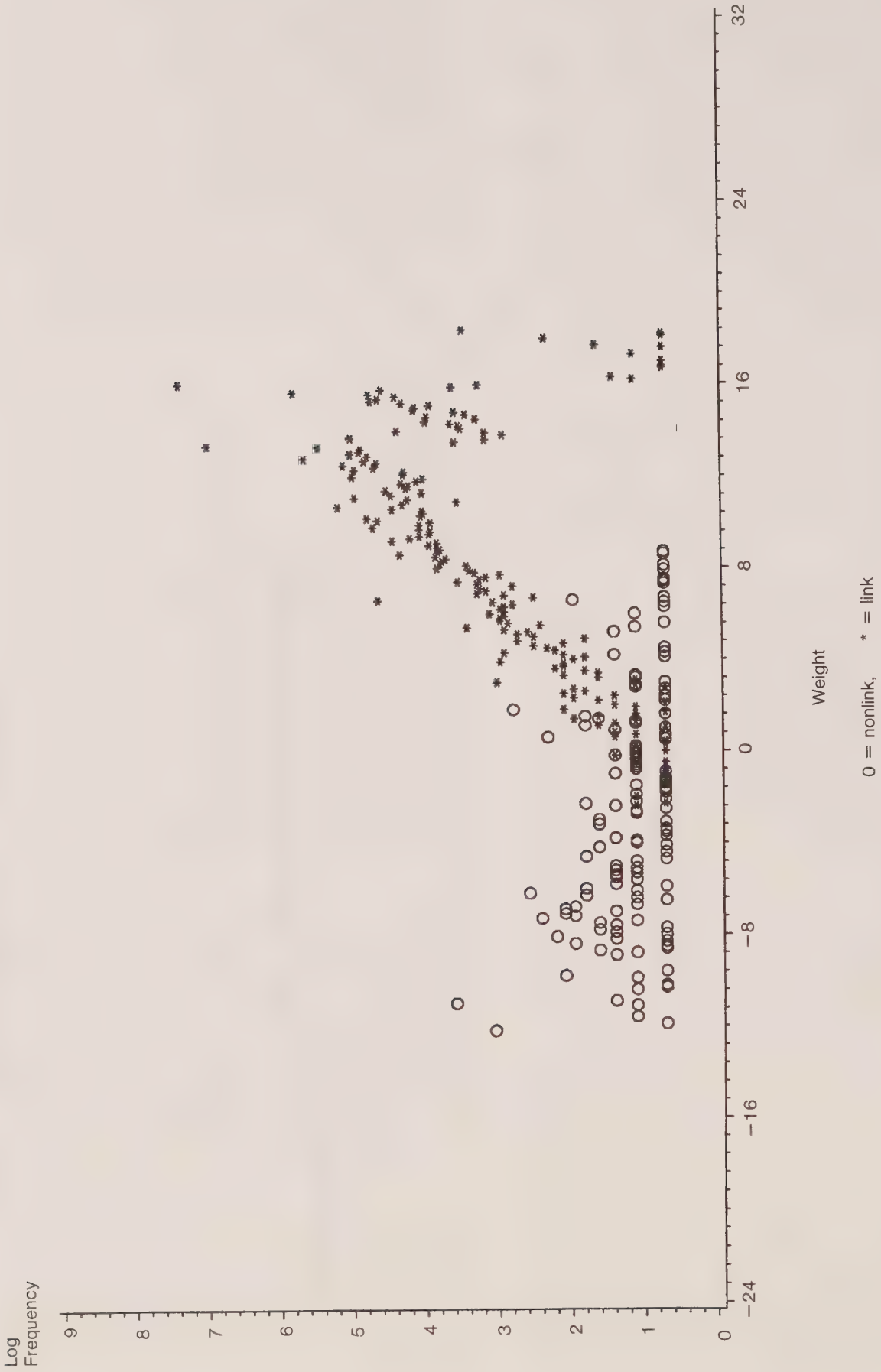


Figure 3. Log of Frequency vs Weight Mediocre Matching Scenario, Links and Nonlinks

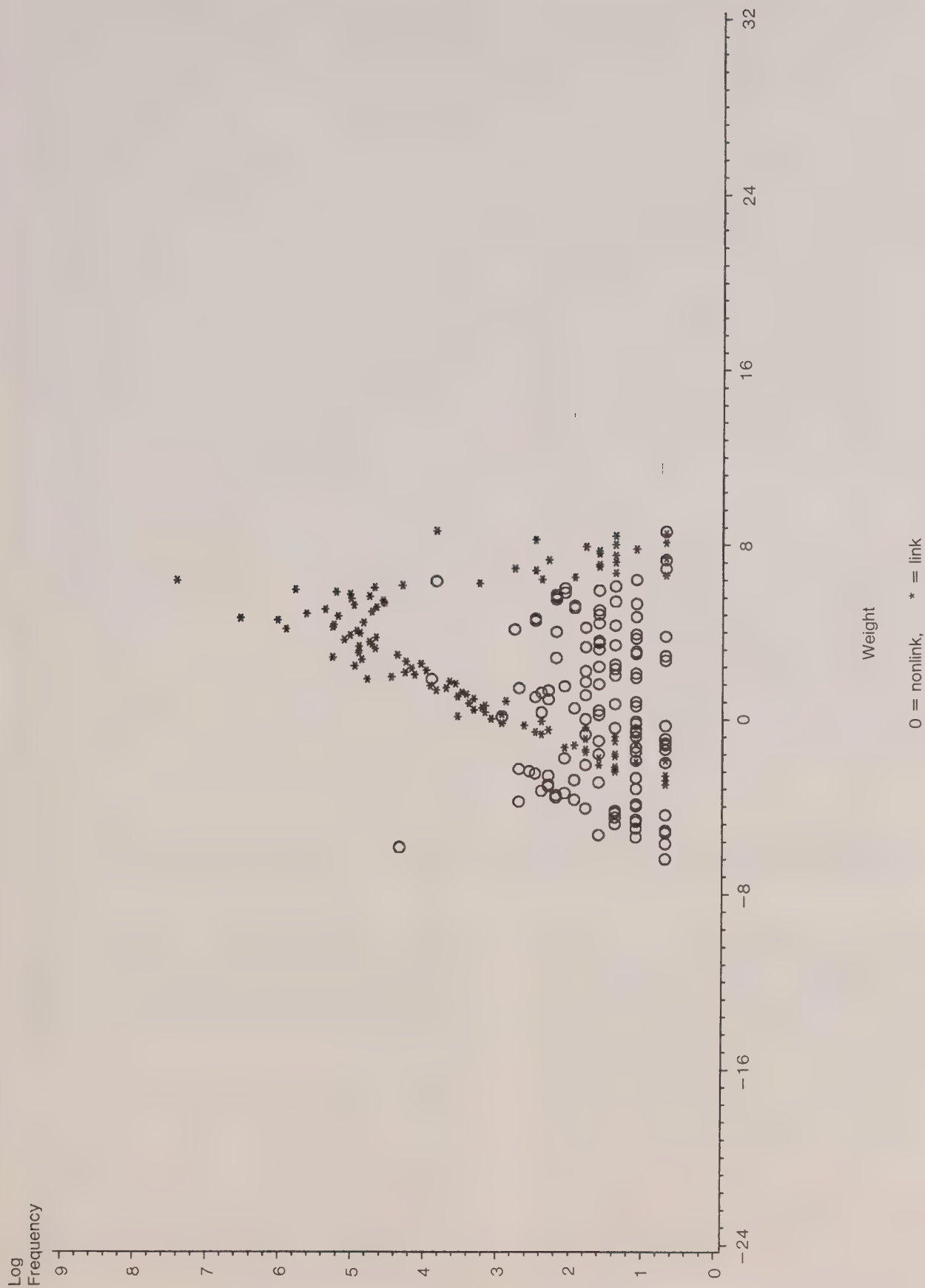


Figure 4. Log of Frequency vs Weight Poor Matching Scenario, Links and Nonlinks

Table 1

Counts of True Links and True Nonlinks and Probabilities of an Erroneous Link in Weight Ranges for Various Matching Cases; Estimated Probabilities via Rubin-Belin Methodology

Weight	False match rates											
	Good				Mediocre				Poor			
	True		Prob		True		Prob		True		Prob	
	Link	NL	True	Est	Link	NL	True	Est	Link	NL	True	Est
15 +	9,176	0	.00	.00	2,621	0	.00	.00	0	1	.00	.00
14	111	0	.00	.00	418	0	.00	.00	0	1	.00	.00
13	91	0	.00	.01	1,877	0	.00	.00	0	1	.00	.00
12	69	0	.00	.02	1,202	0	.00	.00	0	1	.00	.00
11	59	0	.00	.03	832	0	.00	.00	0	1	.00	.00
10	69	0	.00	.05	785	0	.00	.00	0	1	.00	.00
9	42	0	.00	.08	610	0	.00	.00	0	1	.00	.00
8	36	2	.05	.13	439	3	.00	.00	65	1	.02	.00
7	30	1	.03	.20	250	4	.00	.01	39	1	.03	.00
6	14	7	.33	.29	265	9	.03	.03	1,859	57	.03	.03
5	28	4	.12	.40	167	8	.05	.06	1,638	56	.03	.03
4	6	3	.33	.51	89	6	.06	.11	2,664	62	.02	.05
3	12	7	.37	.61	84	5	.06	.20	1,334	31	.02	.11
2	8	6	.43	.70	38	7	.16	.31	947	30	.03	.19
1	7	13	.65	.78	33	34	.51	.46	516	114	.18	.25
0	7	4	.36	.83	13	19	.59	.61	258	65	.20	.28
-1	3	5	.62	.89	7	20	.74	.74	93	23	.20	.31
-2	0	11	.99	.91	3	11	.79	.84	38	23	.38	.41
-3	4	6	.60	.94	4	19	.83	.89	15	69	.82	.60
-4	4	3	.43	.95	0	15	.99	.94	1	70	.99	.70
-5	4	4	.50	.97	0	15	.99	.96	0	25	.99	.68
-6	0	5	.99	.98	0	27	.99	.98	0	85	.99	.67
-7	1	6	.86	.98	0	40	.99	.99			.99	.99
-8	0	8	.99	.99	0	41		.99			.99	.99
-9	0	4	.99	.99	0	4		.99			.99	.99
-10 -	0	22			0	22		.99			.99	.99

Notes: In the first column, weight 10 means weight range from 10 to 11. Weight ranges 15 and above and weight ranges -9 and below are added together. Weights are log ratios that are based on estimated agreement probabilities. **NL** is nonlinks and **Prob** is probability.

parameters given by equation (2.1). The poor matching scenario can arise when we do not have suitable name parsing software that allows comparison of corresponding surnames and first names or suitable address parsing software that allows comparison of corresponding house numbers and street names. Lack of proper parsing means that corresponding matching variables associated with many true links will not be properly utilized.

Our ability to estimate the probability of a match varies significantly. In Table 1 we have displayed these probabilities, both true and estimated, by weight classes. For the good and mediocre matching scenarios, estimated probabilities were fairly close to the true values. For the poor scenario, in which most pairs are potential links, deviations are quite substantial.

For each matching scenario, empirical data were created. Each data base contained a computer matching weight, true and estimated matching probabilities, the independent x -variable for the regression, the true dependent y -variable, the observed y -variables in the record having the highest match weight, and the observed y -variable from the record having the second highest matching weight.

The independent x -variables for the regression were constructed using the SAS RANUNI procedure, so as to be uniformly distributed between 1 and 101. For this paper, they were chosen independently of any matching variables. (While we have considered the situation for which regression variables are dependent on one or more matching variables (Winkler and Scheuren 1991), we do not present any such results in this paper.)

Three regression scenarios were then considered. They correspond to progressively lower R^2 values: (1) R^2 between 0.75 and 0.80; (2) between 0.40 and 0.45; and (3) between 0.20 and 0.22. The dependent variables were generated with independent seeds using the SAS RANNOR procedure. Within each matching scenario (good, mediocre, or poor), all pairing of records obtained by the matching process and, thus, matching error was fixed.

It should be noted that there are two reasons why we generated the (x,y) -data used in the analyses. First, we wanted to be able to control the regression data sufficiently well to determine what the effect of matching error was. This was an important consideration in the very large Monte Carlo simulations reported in Winkler and Scheuren (1991). Second, there existed no available pairs of data files in which highly precise matching information is available and which contain suitable quantitative data.

In performing the simulations for our investigation, some of which are reported here, we created more than 900 data bases, corresponding to a large number of variants of the three basic matching scenarios. Each data base contained three pairs of (x,y) -variables corresponding to the three basic regression scenarios. An examination of these data bases was undertaken to look at some of the matching sensitivity of the regressions and associated adjustments to the sampling procedure. The different data bases determined by different seed numbers are called *different samples*.

The regression adjustments were made separately for each weight class shown in Table 1, using both the estimated and true probabilities of linkage. In Table 1, weight class 10 refers to pairs having weights between 10 and 11 and weight class -1 refers to pairs having weights between -0 and -1 . All pairs having weights 15 and above are combined into class 15+ and all pairs having weights -9 and below are combined into class $-10-$. While it was possible with the Rubin-Belin results to make individual adjustments for linkage probabilities, we chose to make average adjustments, by each weight class in Table 1. (See Czajka *et al.* 1992, for discussion of a related decision. Our approach has some of the flavor of the work on propensity scores (*e.g.*, Rosenbaum and Rubin 1983, 1985). Propensity scoring techniques, while proposed for other classes of problems, may have application here as well.

4. SOME HIGHLIGHTS AND LIMITATIONS OF THE SIMULATION RESULTS

Because of space limitations, we will present only a few representative results from the simulations conducted. For more information, including an extensive set of tables, see Winkler and Scheuren (1991).

The two outcome measures from our simulation that we consider are the relative bias and relative standard

error. We will only discuss the mediocre matching scenario in detail and only for the case R^2 between 0.40 and 0.45. Figures 5-7 shows the relative bias results from a single representative sample. An overall summary, though, for the other scenarios is presented in Table 2. Some limitations on the simulation are also noted at the end of this section.

4.1 Illustrative Results for Mediocre Matching

Rather than use all pairs, we only consider pairs having weights 10 or less. Use of the smaller subset of pairs allows us to examine regression adjustment procedures for weight classes having low to high proportions of true nonlinks. We note that the eliminated pairs (having weight 10 and above) are associated only with true links. Figures 5 and 6 present our results for adjusted and unadjusted regression data, respectively. Results obtained with unadjusted data are based on conventional regression formulas (*e.g.*, Draper and Smith 1981). The weight classes displayed are cumulative beginning with pairs having the highest weight. Weight class w refers to all pairs having weights between w and 10.

We observe the following:

- The *accumulation* is by decreasing matching weight (*i.e.* from classes most likely to consist almost solely of true links to the classes containing increasing higher proportions of true nonlinks). In particular, for weight class $w = 8$, the first data point shown in Figures 5-7, there were 3 nonlinks and 439 links. By the time, say, we had cumulated the data through weight class $w = 5$, there were 24 nonlinks; the links, however, had grown to 1,121 – affording us a much larger overall sample size with a corresponding reduction in the regression standard error.
- Relative *biases* are provided for the original and adjusted slope coefficient \hat{a}_1 by taking the ratio of the true coefficient (about 2) and the calculated one for each cumulative weight class.
- Adjusted regression results are shown employing both estimated and true match probabilities. In particular, Figure 5 corresponds to the results obtained using estimated probabilities (all that would ordinarily be available in practice). Figure 7 corresponds to the unrealistic situation for which we knew the true probabilities.
- Relative *root mean square errors* (not shown) are obtained by calculating MSEs for each cumulative weight class. For each class, the bias is squared, added to the square of the standard errors, and square roots taken.

Observations on the results we obtained are fairly straightforward and about what we expected. For example, as sample size increased, we found the relative root mean square errors decreased substantially for the adjusted coefficients. If the regression coefficients were not adjusted,

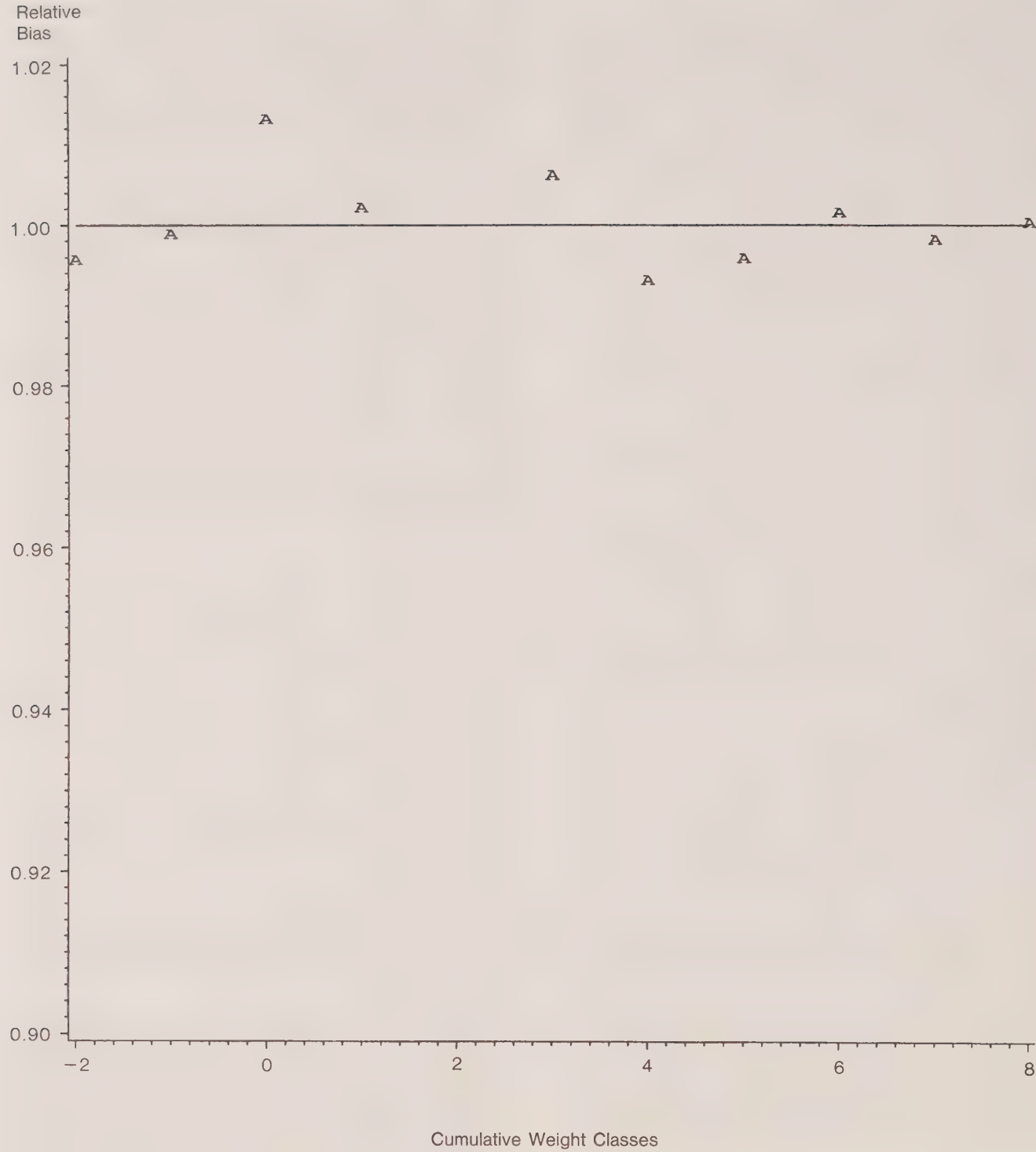


Figure 5. Relative Bias For Adjusted Estimators, Estimated Probabilities

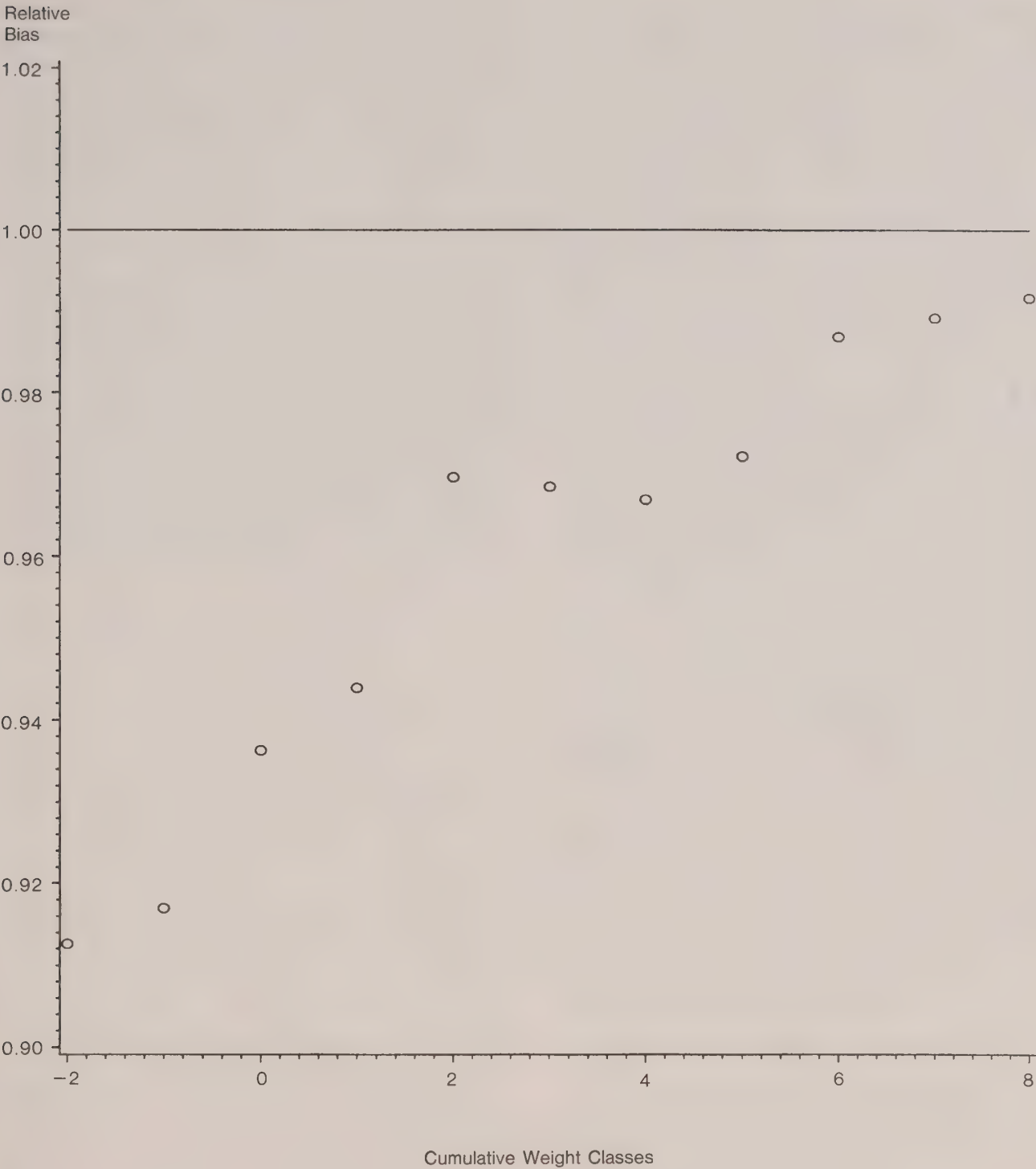


Figure 6. Relative Bias For Unadjusted Estimators

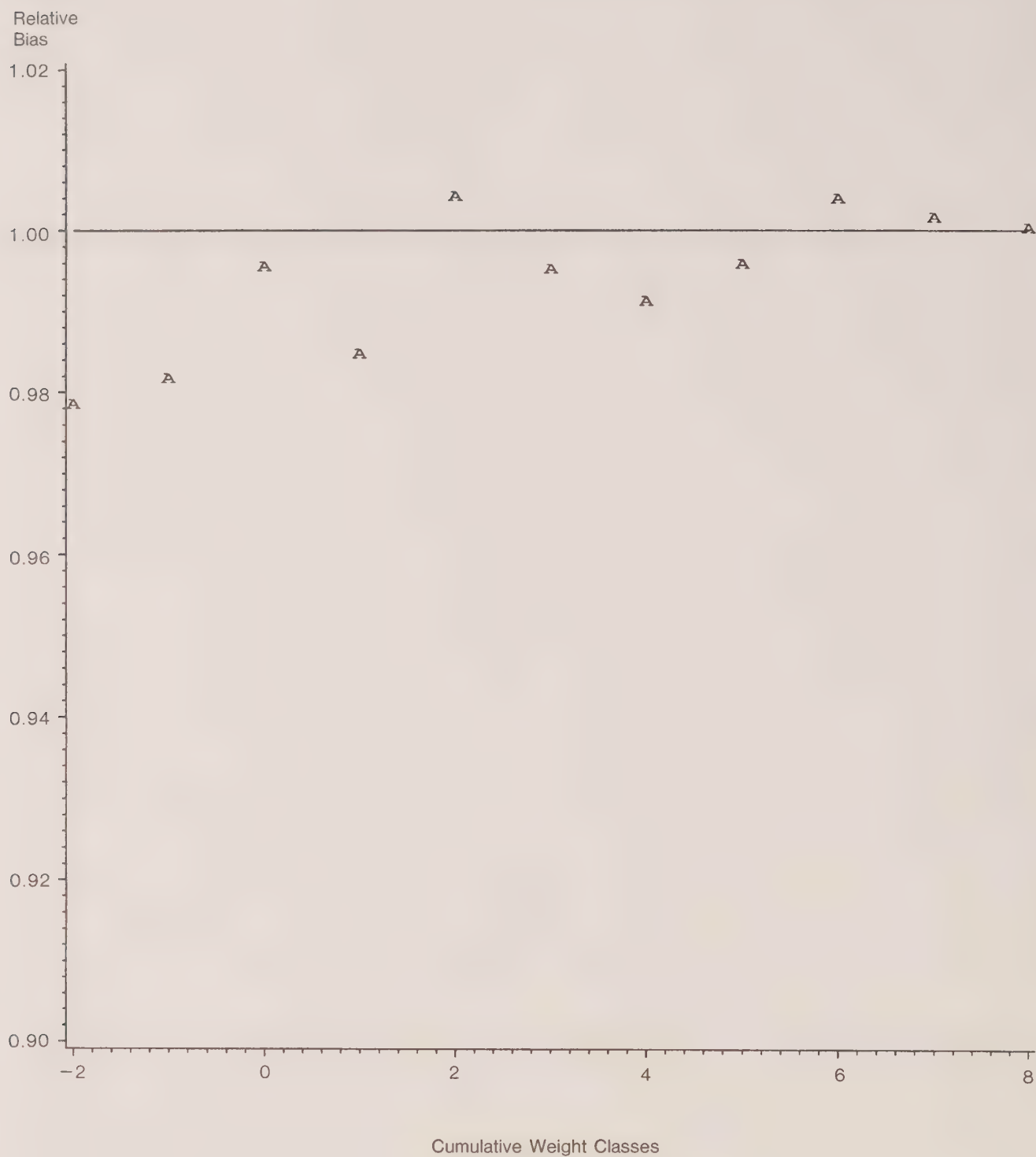


Figure 7. Relative Bias For Adjusted Estimators, True Probabilities

standard errors still decreased as the sample size grew, but at an unacceptably high price in increased bias.

One point of concern is that our ability to accurately estimate matching probabilities critically affects the accuracy of the coefficient estimates. If we can accurately estimate the probabilities (as in this case), then the adjustment procedure works reasonably well; if we cannot (see below), then the adjustment could perform badly.

4.2 Overall Results Summary

Our results varied somewhat for the three different values of R^2 – being better for larger R^2 values. These R^2 differences, however, do not change our main conclusions; hence, Table 2 does not address them. Notice that, for the good matching scenario, attempting to adjust does little good and may even cause some minor harm. Certainly it is pointless, in any case, and we only included it in our simulations for the sake of completeness. At the other extreme, even for poor matches, we obtained satisfactory results, but only when using the true probabilities – something not possible in practice.

Table 2

Summary of Adjustment Results for
Illustrative Simulations

Basis of adjustments	Matching scenarios		
	Good	Mediocre	Poor
True probabilities	Adjustment was not helpful because it was not needed	Good results like those in Section 4.1	Good results like those in Section 4.1
Estimated probabilities	Same as above	Same as above	Poor results because Rubin-Belin could not estimate the probabilities

Any statistical estimation procedure will have difficulty with the poor matching scenario because of the extreme overlap of the curves. See Figure 4. We believe the mediocre scenario covers a wide range of typical settings. Nonetheless, the poor matching scenario might arise fairly often too, especially with less experienced linkers. Either new estimation procedures will have to be developed for the poor case or the Rubin-Belin probability estimation procedure – which was not designed for this situation – will have to be enhanced.

4.3 Some Simulation Limitations

The simulation results are subject to a number of limitations. Some of these are of possible major practical significance; others less so. A partial list follows:

- In conducting simulations for this paper, we assumed that the highest weight pair was a true link and the second highest a true nonlink. This assumption fails because, sometimes, the second highest is the true link and the highest a true nonlink. (We do not have a clear sense of how important this issue might be in practice. It would certainly have to be a factor in poor matching scenarios.)
- A second limitation of the data sets employed for the simulations is that the truly linked record may not be present at all in the file to which the first file is being matched. (This could be important. In many practical settings, we would expect the “logical blocking criteria” also to cause both pairs used in the adjustment to be false links.)
- A third limitation of our approach is that no use has been made of conventional regression diagnostic tools. (Depending on the environment, outliers created because of nonlinks could wreak havoc with underlying relationships. In our simulations this did not show up as much of a problem, largely, perhaps, because the X and Y values generated were bounded in a moderately narrow range.)

5. CONCLUSIONS AND FUTURE WORK

The theoretical and related simulation results presented here are obviously somewhat contrived and artificial. A lot more needs to be done, therefore, to validate and generalize our beginning efforts. Nonetheless, some recommendations for current practice stand out, as well as areas for future research. We will cover first a few of the topics that intrigued us as worthy of more study to improve the adjustment of potential links. Second, some remarks are made about the related problem of what to do with the (remaining) nonlinks. Finally, the section ends with some summary ideas and a revisitation of our perspective concerning the unity of the tasks that linkers and analysts do.

5.1 Improvements in Linkage Adjustment

An obvious question is whether our adjustment procedures could borrow ideas from general methods for errors-in-variables (*e.g.*, Johnston 1972). We have not explored this, but there may be some payoffs.

Of more interest to us are techniques that grow out of conventional regression diagnostics. A blend of these with our approach has a lot of appeal. Remember we are making adjustments, weight class by weight class. Suppose we looked ahead of time at the residual scatter in a particular weight class, where the residuals were calculated around the regression obtained from the cumulative weight classes above the class in question. Outliers, say, could then be identified and might be treated as nonlinks rather than potential links.

We intend to explore this possibility with simulated data that is heavier-tailed than what was used here. Also we will explore consciously varying the length of the weight classes and the minimum number of cases in each class. We have an uneasy feeling that the number of cases in each class may have been too small in places. (See Table 1.) On the other hand, we did not use the fact that the weight classes were of equal length nor did we study what would have happened had they been of differing lengths.

One final point, as noted already: we believe our approach has much in common with propensity scoring, but we did not explicitly appeal to that more general theory for aid and this could be something worth doing. For example, propensity scoring ideas may be especially helpful in the case where the regression variables and the linkage variables are dependent. (See Winkler and Scheuren (1991) for a report on the limited simulations undertaken and the additional difficulties encountered.)

5.2 Handling Erroneous Nonlinks

In the use of record linkage methods the general problem of selection bias arises because of erroneous nonlinks. There are a number of ways to handle this. For example, the links could be adjusted by the analyst for lack of representativeness, using the approaches familiar to those who adjust for unit or, conceivably, item nonresponse (e.g., Scheuren *et al.* 1981).

The present approach for handling potential links could help reduce the size of the erroneous nonlink problem but, generally, would not eliminate it. To be specific, suppose we had a linkage setting where, for resource reasons, it was infeasible to follow up on the potential links. Many practitioners might simply drop the potential links, thereby, increasing the number of erroneous nonlinks. (For instance, in ascertaining which of a cohort's members is alive or dead, a third possibility – unascertained – is often used.)

Our approach to the potential links would have *implicitly* adjusted for that portion of the erroneous nonlinks which were potentially linkable (with a followup step, say). Other erroneous nonlinks would generally remain and another adjustment for them might still be an issue to consider.

Often we can be faced with linkage settings where the files being linked have subgroups with matching information of varying quality, resulting in differing rates of erroneous links and nonlinks. In principle, we could employ the techniques in this paper to each subgroup separately. How to handle very small subgroups is an open problem and the effect on estimated differences between subgroups, even when both are of modest size, while seemingly straightforward, deserves study.

5.3 Concluding Comments

At the start of this paper we asked two “key” questions. Now that we are concluding, it might make sense to reconsider

these questions and try, in summary fashion, to give some answers.

- “*What should the linker do to help the analyst?*” If possible, the linker should play a role in designing the datasets to be matched, so that the identifying information on both is of high quality. Powerful algorithms exist now in several places to do an excellent job of linkage (e.g., at Statistics Canada or the U.S. Bureau of the Census, to name two). Linkers should resist the temptation to design and develop their own software. In most cases, modifying or simply using existing software is highly recommended (Scheuren 1985). Obviously, for the analyst's sake, the linker needs to provide as much linkage information as possible on the files matched so that the analyst can make informed choices in his or her work. In the present paper we have proposed that the links, nonlinks, and potential links be provided to the analyst – not just links. We strongly recommend this, even if a clerical review step has been undertaken. We do *not* necessarily recommend the particular choices we made about the file structure, at least not without further study. We would argue, though, that our choices are serviceable.

- “*What should the analyst know about the linkage and how should this be used?*” The analyst needs to have information like link, nonlink, and potential link status, along with linkage probabilities, if available. Many settings could arise where simply doing the data analysis steps separately by link status will reveal a great deal about the sensitivity of one's results. The present paper provides some initial ideas about how this use might be approached in a regression context. There also appears to be some improvements possible using the adjustments carried out here, particularly for the mediocre matching scenario. How general these improvements are remains to be seen. Even so, we are relatively pleased with our results and look forward to doing more. Indeed, there are direct connections to be made between our approach to the regression problem and other standard techniques, like contingency table loglinear models.

Clearly, we have not developed complete, general answers to the questions we raised. We hope, though, that this paper will at least stimulate interest on the part of others that could lead us all to better practice.

ACKNOWLEDGMENTS AND DISCLAIMERS

The authors would like to thank Yahia Ahmed and Mary Batcher for their help in preparing this paper and two referees for detailed and discerning comments. Fruitful discussions were held with Tom Belin. Wendy Alvey also provided considerable editorial assistance.

The usual disclaimers are appropriate here: in particular, this paper reflects the views of the authors and not necessarily those of their respective agencies. Problems, like a lack of clarity in our thinking or in our exposition, are entirely the authors' responsibility.

APPENDIX

The appendix is divided into four sections. The first provides details on how matching error affects regression models for the simple univariate case. The approach most closely resembles the approach introduced by Neter *et al.* (1965) and provides motivation for the generalizations presented in appendix sections two and three. Computational formulas are considerably more complicated than those presented by Neter *et al.* because we use a more realistic model of the matching process. In the second section, we extend the univariate model to the case for which all independent variables arise from one file, while the dependent variable comes from the other, and, in the third, we extend the second case to that in which some independent variables come from one file and some come from another. The fourth section summarizes methods of Rubin and Belin (1991) (see also Belin 1991) for estimating the probability of a link.

A.1. Univariate Regression Model

In this section we address the simplest regression situation in which we match two files and consider a set of numeric pairs in which the independent variable is taken from a record in one file and the dependent variable is taken from the corresponding matched record from the other file.

Let $Y = X\beta + \epsilon$ be the ordinary univariate regression model for which error terms are independent with expectation zero and constant variance σ^2 . If we were working with a single data base, Y would be regressed on X in the usual manner. For $i = 1, \dots, n$, we wish to use (X_i, Y_i) but we will use (X_i, Z_i) , where Z_i is usually Y_i but it may take some other value Y_j due to matching error.

That is, for $i = 1, \dots, n$,

$$z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij} \text{ for } j \neq i, \end{cases}$$

where $p_i + \sum_{j \neq i} q_{ij} = 1$.

The probability p_i may be zero or one. We define $h_i = 1 - p_i$. As in Neter *et al.* (1965), we divide the set of pairs into n mutually exclusive classes. Each class consists of exactly one (X_i, Z_i) and, thus, there are n classes. The intuitive idea of our procedure is that we basically adjust

Z_i in each (X_i, Z_i) for the bias induced by the matching process. The accuracy of the adjustment is heavily dependent on the accuracy of the estimates of the matching probabilities in our model.

To simplify the computational formulas in the explanation, we assume one-to-one matching; that is, for each $i = 1, \dots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$. Our model still applies if we do not assume one-to-one matching.

As intermediate steps in estimating regression coefficients and their standard errors, we need to find $\mu_z \equiv E(Z)$, σ_z^2 , and σ_{zx} . As in Neter *et al.* (1965),

$$\begin{aligned} E(Z) &\equiv (1/n) \sum_i E(Z|i) \equiv (1/n) \sum_i (Y_i p_i + \sum_{j \neq i} Y_j q_{ij}) \\ &= (1/n) \sum_i Y_i \\ &\quad + (1/n) \sum_i [Y_i (-h_i) + Y_{\phi(i)} h_i] \\ &\equiv \bar{Y} + B. \end{aligned} \quad (\text{A.1.1})$$

The first and second equalities are by definition and the third is by addition and subtraction. The third inequality is the first time we apply the one-to-one matching assumption. The last term on the right hand side of the equality is the bias which we denote by B . Note that the overall bias B is the statistical average (expectation) of the individual biases $[Y_i (-h_i) + Y_{\phi(i)} h_i]$ for $i = 1, \dots, n$. Similarly, we have

$$\begin{aligned} \sigma_z^2 &\equiv E(Z - EZ)^2 = E(Z - (\bar{Y} + B))^2 \\ &= (1/n) \sum_i (Y_i - \bar{Y})^2 p_i + (1/n) \sum_{j \neq i} \\ &\quad (Y_j - \bar{Y})^2 q_{ij} - 2B E(Z - \bar{Y}) + B^2 \\ &= (1/n) S_{yy} + B_{yy} - B^2 = \sigma_y^2 + B_{yy} - B^2, \end{aligned} \quad (\text{A.1.2})$$

where $B_{yy} = (1/n) \sum_i [(Y_i - \bar{Y})^2 (-h_i) + (Y_{\phi(i)} - \bar{Y})^2 h_i]$, $S_{yy} = \sum_i (Y_i - \bar{Y})^2$ and $\sigma_y^2 = (1/n) S_{yy}$.

$$\begin{aligned} \sigma_{zx} &\equiv E[(Z - EZ)(X - EX)] \\ &= (1/n) \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) p_i \\ &\quad + (1/n) \sum_{j \neq i} (Y_j - \bar{Y})(X_i - \bar{X}) q_{ij} \\ &= (1/n) S_{yx} + B_{yx} = \sigma_{yx} + B_{yx}, \end{aligned} \quad (\text{A.1.3})$$

where $B_{yx} = (1/n) \sum_i [(Y_i - \bar{Y})(X_i - \bar{X})(-h_i) + (Y_{\phi(i)} - \bar{Y})(X_i - \bar{X})h_i]$, $S_{yx} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$, and $\sigma_{yx} = (1/n)S_{yx}$. The term B_{yy} is the bias for the second moments and the term B_{yx} is the bias for the cross-product of Y and X . Formulas (A.1.1), (A.1.2), and (A.1.3), respectively, correspond to formulas (A.1), (A.2), and (A.3) in Neter *et al.* The formulas necessarily differ in detail because we use a more general model of the matching process.

The regression coefficients are related by

$$\beta_{zx} \equiv \sigma_{zx}/\sigma_x^2 = \sigma_{yx}/\sigma_x^2 + B_{yx}/\sigma_x^2 = \beta_{yx} + B_{yx}/\sigma_x^2. \quad (\text{A.1.4})$$

To get an estimate of the variance of β_{yx} , we first derive an estimate s^2 for the variance σ^2 in the usual manner.

$$\begin{aligned} (n-2)s^2 &= \sum_i (y_i - \hat{y}_i)^2 = S_{yy} + \beta_{yx} S_{yx} \\ &= n\sigma_y^2 - n\beta_{yx}\sigma_x^2. \end{aligned} \quad (\text{A.1.5})$$

Using (A.1.2) and (A.1.3) allows us to express s^2 in terms of the observable quantities σ_z^2 and σ_{zx} and the bias terms B_{yy} , B_{yx} , and B that are computable under our assumptions. The estimated variance of β_{yx} is then computed by the usual formula (e.g., Draper and Smith 1981, 18-20)

$$\text{Var}(\beta_{yx}) = s^2/(n\sigma_x^2).$$

We observe that the first equality in (A.1.5) involves the usual regression assumption that the error terms are independent with identical variance.

In the numeric examples of this paper we assumed that the true independent value X_i associated with each Y_i was from the record with the highest matching weight and the false independent value was taken from the record with the second highest matching weight. This assumption is plausible because we have only addressed simple regression in this paper and because the second highest matching weight was typically much lower than the highest. Thus, it is much more natural to assume that the record with the second highest matching weight is false. In our empirical examples we use straightforward adjustments and make simplistic assumptions that work well because they are consistent with the data and the matching process. In more complicated regression situations or with other models such as loglinear we will likely have to make additional modelling assumptions. The additional assumptions can be likened to the manner in which simple models for nonresponse require additional assumptions as the models progress from ignorable to nonignorable (see Rubin 1987).

In this section, we chose to adjust independent x -values and leave dependent y -values as fixed in order to achieve consistency with the reasoning of Neter *et al.* We could have just as easily adjusted dependent y -values leaving x -values as fixed.

A.2. Multiple Regression with Independent Variables from One File and Dependent Variables from the Other File

At this point we pass to the usual matrix notation (e.g., Graybill 1976). Our basic model is

$$Y = X\beta + \epsilon,$$

where Y is a $n \times 1$ array, X is a $n \times p$ array, β is a $p \times 1$ array, and ϵ is a $n \times 1$ array.

Analogous to the reasoning we used in (A.1.1), we can represent

$$Z = Y + B, \quad (\text{A.2.1})$$

where Z , Y , and B are $n \times 1$ arrays having terms that correspond, for $i = 1, \dots, n$, via

$$z_i = y_i + p_i y_i + h_i y_{\phi(i)}.$$

Because we observe Z and X only, we consider the equation

$$Z = XC + \epsilon. \quad (\text{A.2.2})$$

We obtain an estimate \hat{C} by regressing on the observed data in the usual manner. We wish to adjust the estimate \hat{C} to an estimate $\hat{\beta}$ of β in a manner analogous to (A.1.1).

Using (A.2.1) and (A.2.2) we obtain

$$(X^T X)^{-1} X^T Y + (X^T X)^{-1} X^T B = \hat{C}. \quad (\text{A.2.3})$$

The first term on the left hand side of (A.2.3) is the usual estimate $\hat{\beta}$. The second term on the left hand side of (A.2.3) is our bias adjustment. X^T is the transpose of X .

The usual formula (Graybill 1976, p. 176) allows estimation of the variance σ^2 associated with the i.i.d. error components of ϵ ,

$$\begin{aligned} (n-p)\hat{\sigma}^2 &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= Y^T Y - \hat{\beta}^T X^T Y, \end{aligned} \quad (\text{A.2.4})$$

where $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Via (A.2.1) $\hat{\beta}^T X^T Y$ can be represented in terms of the observable Z and X in a manner similar to (A.1.2) and (A.1.3). As

$$Y^T Y = Z^T Z - B^T Z - Z^T B + B^T B, \quad (\text{A.2.5})$$

we can obtain the remaining portion of the right hand side of (A.2.4) that allows estimation of σ^2 .

Via the usual formula (e.g., Graybill 1976, p. 276), the covariance of $\hat{\beta}$ is

$$\text{cov}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}, \quad (\text{A.2.6})$$

which we can estimate.

A.3. Multiple Regression with Independent Variables from Both Files

When some of the independent variables come from the same file as Y we must adjust them in a manner similar to the way in which we adjust Y in equations (A.1.1) and (A.2.1). Then data array X can be written in the form

$$X_d = X + D, \quad (\text{A.3.1})$$

where D is the array of bias adjustments taking those terms of X arising from the same file as Y back to their true values that are represented in X_d . Using (A.2.1) and (A.2.2), we obtain

$$Y + B = (X_d - D)C. \quad (\text{A.3.2})$$

With algebra (A.3.2) becomes

$$\begin{aligned} (X_d^T X_d)^{-1} X_d^T Y &= (X_d^T X_d)^{-1} X_d^T (-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T (X_d + D)C \\ &= (X_d^T X_d)^{-1} X_d^T (-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T DC + C. \end{aligned} \quad (\text{A.3.3})$$

If D is zero (*i.e.*, all independent x -values arise from a single file), then (A.3.3) agrees with (A.2.3). The first term on the left hand side of (A.2.3) is the estimate of $\hat{\beta}$. The estimate $\hat{\sigma}^2$ is obtained analogously to the way (A.2.3), (A.2.4) and (A.2.5) were used. The covariance of $\hat{\beta}$ follows from (A.2.6).

A.4. Rubin-Belin Model

To estimate the probability of a true link within any weight range, Rubin and Belin (1991) consider the set of pairs that are produced by the computer matching program and that are ranked by decreasing weight. They assume that the probability of a true link is a monotone function of the weight; that is, the higher the weight, the higher the probability of a true link. They assume that the distribution of the observed weights is a mixture of the distributions for true links and true nonlinks.

Their estimation procedure is:

1. Model each of the two components of the mixture as normal with unknown mean and variance after separate power transformations.
2. Estimate the power of the two transformations from a training sample.
3. Taking the two transformations as known, fit a normal mixture model to the current weight data to obtain maximum likelihood estimates (and standard errors).

4. Use the parameters from the fitted model to obtain point estimates of the false-link rate as a function of cutoff level and obtain standard errors for the false-link rate using the delta-method approximation.

While the Rubin-Belin method requires a training sample, the training sample is primarily used to get the shape of the curves. That is, if the power transformation is given by

$$\psi(w_i; \delta, \omega) = \begin{cases} (w_i^\delta - 1) / (\delta \omega^{\delta-1}) & \text{if } \delta \neq 0 \\ \omega \log(w_i) & \text{if } \delta = 0, \end{cases}$$

where ω is the geometric mean of the weights w_i , $i = 1, \dots, n$, then ω and δ can be estimated for the two curves. For the examples of this paper and a large class of other matching situations (Winkler and Thibaudeau 1991), the Rubin-Belin estimation procedure works well. In some other situations a different method (Winkler 1992) that uses more information than the Rubin-Belin method and does not require a training sample yields accurate estimates, while software (see *e.g.*, Belin 1991) based on the Rubin-Belin method fails to converge even if new calibration data are obtained. Because the calibration data for the good and mediocre scenarios of this paper are appropriate, the Rubin-Belin method provides better estimates than the method of Winkler.

REFERENCES

- BEEBE, G. W. (1985). Why are epidemiologists interested in matching algorithms? In *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service.
- BELIN, T. (1991). Using Mixture Models to Calibrate Error Rates in Record Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation. Harvard Ph.D. Thesis.
- CARPENTER, M., and FAIR, M.E. (Editors) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, Statistics Canada.
- COOMBS, J.W., and SINGH, M.P. (Editors) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.
- COPAS, J.B., and HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society A*, 153, 287-320.
- CZJAKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, D.B. (1992). Evaluation of a new procedure for estimating income and tax aggregates from advance data. *Journal of Business and Economic Statistics*, 10, 117-131.
- DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis*, 2nd Edition. New York: J. Wiley.

- FELLEGI, I.P., and SUNTER, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.
- HOWE, G., and SPASOFF, R.A. (Editors) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto, Ontario, Canada: University of Toronto Press.
- JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.
- JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- JOHNSTON, J. (1972). *Econometric Methods*, 2nd Edition. New York: McGraw-Hill.
- KILSS, B., and ALVEY, W. (Editors) (1985). *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service, Publication 1299, 2-86.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford: Oxford University Press.
- NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1208.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NETER, J., MAYNES, E.S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- ROSENBAUM, P., and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- ROSENBAUM, P., and RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.
- RUBIN, D.B. (1990). Discussion (of Imputation Session). *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, 676-678.
- RUBIN, D., and BELIN, T. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.
- SCHEUREN, F. (1985). Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service.
- SCHEUREN, F., and OH, H.L. (1975). Fiddling Around with Nonmatches and Mismatches. *Proceedings of the Social Statistics Section, American Statistical Association*, 627-633.
- SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. *Studies from Interagency Data Linkages*, U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.
- TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- WINKLER, W.E. (1985). Exact matching list of businesses: blocking, subfield identification, and information theory. In *Record Linkage Techniques - 1985*, (Eds. B. Kilss and W. Alvey). U.S. Internal Revenue Service, Publication 1299, 2-86.
- WINKLER, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.
- WINKLER, W.E., and SCHEUREN, F. (1991). How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis. U.S. Bureau of the Census, Statistical Research Division Technical Report.
- WINKLER, W.E., and THIBAUDEAU, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical Report.

Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption

A.C. SINGH, H.J. MANTEL, M.D. KINACK and G. ROWE¹

ABSTRACT

In the creation of micro-simulation databases which are frequently used by policy analysts and planners, several datafiles are combined by statistical matching techniques for enriching the host datafile. This process requires the conditional independence assumption (CIA) which could lead to serious bias in the resulting joint relationships among variables. Appropriate auxiliary information could be used to avoid the CIA. In this report, methods of statistical matching corresponding to three methods of imputation, namely, regression, hot deck, and log linear, with and without auxiliary information are considered. The log linear methods consist of adding categorical constraints to either the regression or hot deck methods. Based on an extensive simulation study with synthetic data, sensitivity analyses for departures from the CIA are performed and gains from using auxiliary information are discussed. Different scenarios for the underlying distribution and relationships, such as symmetric versus skewed data and proxy versus nonproxy auxiliary data, are created using synthetic data. Some recommendations on the use of statistical matching methods are also made. Specifically, it was confirmed that the CIA could be a serious limitation which could be overcome by the use of appropriate auxiliary information. Hot deck methods were found to be generally preferable to regression methods. Also, when auxiliary information is available, log linear categorical constraints can improve performance of hot deck methods. This study was motivated by concerns about the use of the CIA in the construction of the Social Policy Simulation Database at Statistics Canada.

KEY WORDS: Categorical constraints; Conditional correlation; Log normal contaminations; Shrinkage to the mean.

1. INTRODUCTION

Statistical matching can be viewed as a special case of imputation in which we have two distinct micro-data sources containing different information on different units. One data source serves as a host or recipient file to which new information is imputed for each record using data from the other source which is the donor file. Statistical matching, however, differs from the usual problem of imputation whenever the host file contains information about additional variables which are not present in the donor file. A typical use for the matched file is as input to micro-simulation models for which a complete file with all variables is required. Available micro-datafiles may correspond to samples from administrative files or survey data. Since the records from the different source files correspond to different units, the process of merging the information from the various files is unlike exact matching in which one would search through these other data sources for specific units. In fact, even if exact matching were possible, confidentiality concerns could prevent an exact matching of the files.

A general formulation is as follows. A host file A will contain information on variables (X, Y) and a donor file B

will contain information on variables (X, Z) . The common variable X can be used to identify similar units in the two files. The problem is to complete the records in file A by imputing live values for Z , using the information on the (X, Z) relationship in file B. In practice, the variables X, Y , and Z would generally be multivariate. An important advantage of imputing live values of Z is that relationships among components of multivariate Z are preserved. Throughout this paper, it will be assumed, for convenience, that X, Y and Z are univariate.

The Social Policy Simulation Database (SPSD; see Wolfson *et al.* 1987), a micro-simulation database created at Statistics Canada, provides an important application of statistical matching for use in economic policy analysis, *e.g.*, calculations of taxes and transfers for families on the database. The multistage construction process of the SPSP uses the technique of statistical matching at a number of points in order to enrich the host datafile, the Survey of Consumer Finance (SCF), with additional information from other data sources. Specifically, information from unemployment insurance claim histories, personal income tax returns, and the Family Expenditure Survey is added to the SCF records. If file A corresponds to the SCF and file B to the tax file, then X variables may represent

¹ A.C. Singh, H.J. Mantel and M.D. Kinack, Social Survey Methods Division; G. Rowe, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

demographic and economic variables, Y may denote transfer income, and Z may correspond to tax liability, investment income and deductions.

Statistical matching, as described above, suffers from a serious limitation in that information on the variable Y is completely ignored. This limitation amounts to the assumption of conditional independence of Y and Z given X ($Y \perp Z \mid X$), denoted CIA (conditional independence assumption). The importance of the CIA is obvious, since the purpose of the match is to analyze the joint relationships of X , Y and Z . If the true relationships of the variables are such that conditional independence does not hold, then the CIA would mask an important component of these relationships, and would bias some analyses involving the full set of variables. The potential seriousness of the CIA was noted by Sims (1978) and Rubin (1986), and, although statistical matches based on the CIA are not necessarily seriously flawed, Paass (1986) and Armstrong (1989) offer some empirical evidence that the problem is often real. The present study, in fact, is motivated from considerations of improving the content of the SPSPD which assumes the CIA for the process of statistical matching; see also comments of Scheuren (1989) on the methodology used in the SPSPD.

The literature on statistical matching extends over more than two decades. Early references are Budd and Radner (1969), Budd (1971) and Okner (1972). Sims (1972), in his comments on Okner's paper, was the first to point out the potential risk of statistical matching because of the implicit conditional independence assumption. Concerns were also expressed by Fellegi (1977) about the validity of joint distributions in the matched file and he suggested that thorough empirical testing of matching methods should be done. U.S. Department of Commerce (1980) provides a good review of statistical matching as well as exact matching methods; see also Kadane (1978) and Rodgers (1984). Barr and Turner (1990) describe a detailed empirical investigation of quality issues for file merging, and also present a good list of references. For a more recent review see Cohen (1991).

In this paper we consider the use of auxiliary information as an alternative to the CIA in statistical matching. Thus, it is assumed that there exists a third file C representing auxiliary information about the full set (X, Y, Z) or the reduced set (Y, Z) . This information could be outdated, proxy (*i.e.* different but similar variables), or in the form of frequency tables and could come from small scale specially conducted surveys or from confidential datafiles. We wish to complete records in file A by adding Z from file B using information from files A , B , and C on the joint relationships of X , Y , and Z . A measure of success would be the extent to which the Z values on the completed file A could reasonably have come from the true underlying distribution conditional on X and Y . In the context of a simulation study we can compare the matched Z values to

the suppressed true Z values by evaluation measures at the unit level or at the aggregate level. Some examples of unit level measures are mean absolute distance from the true Z values and the deviation of conditional covariance, $\text{Cov}(Y, Z \mid X)$, from the true value. Some examples of aggregate level evaluation measures are chi-square distance and P -values based on likelihood ratio tests for categorical distributions. It is often the case in practice that the completed file A is used to produce cross-classified tables of counts and, therefore, the aggregate level measures based on categorical distributions would generally be of main interest. Moreover, for any arbitrary distribution for (X, Y, Z) , which could be quite complex in practice, the categorical transformation provides a simple unified approach for summarizing the joint distribution.

The statistical matching problem as mentioned above is clearly important from practical considerations. In practice, for a given problem the matching method should be appropriately chosen for the type of auxiliary information available. The methods proposed earlier in the literature are mainly due to Rubin (1986) and Paass (1986). Rubin proposed versions of parametric regression while Paass proposed versions of nonparametric regression. These are related respectively to the familiar regression (REG) and hot deck (HOD) methods of imputation.

Rubin's method (a version of which is denoted in this paper by REG*) basically consists of first finding an intermediate value, Z_{int} , from the regression predictor of Z on X and Y (obtained by using information about the unconditional correlation $\rho_{Y,Z}$ or the conditional correlation $\rho_{Y,Z \mid X}$ from file C) and then a live Z -value is determined from file B using hot deck with (X, Z) Euclidean distance; see Section 3 for details. If the form of the regression predictor function is known, then the REG* procedure for statistical matching could be easily implemented in practice. However, finding a suitable predictor for Z is in general not easy, especially when Z is multivariate. Moreover, if information in file C is in the form of a categorical distribution, which may be quite common in practice, the REG* method would not be applicable.

Paass's method (a version of which is denoted in this paper by HOD*), on the other hand, basically consists of first finding an intermediate value, Z_{int} , from file C by hot deck imputation (with Y - or (X, Y) -distance as the case may be) and then a live Z -value from file B is obtained using again hot deck with (X, Z) -distance. This is a simplified version of the original Paass's method which is iterative such that values of Z for file A , Y for file B , and X for file C (assuming C has only (Y, Z) information) are updated successively using files C , A and B respectively until some convergence criterion is satisfied; see Section 3 for details. To start the iteration, initial values of Z for A , Y for B and X for C are imputed suitably. In the evaluation study considered in this paper, we have considered only the simplified version of Paass's method due to the

considerable computational effort required for the original method. As in the case of the REG* method, the HOD* method is not applicable if file C is in the form of a frequency table. Moreover, even if file C contains micro-data but its size is small (as in the case of a small scale specially conducted survey) or it is proxy or outdated, it may be better to extract some macro-level information such as the categorical distribution based on a fairly coarse partition.

It may be remarked that in the absence of auxiliary information, *i.e.* file C, both REG* and HOD* methods reduce simply to the usual methods of imputation, namely regression (REG) and hot deck (HOD). As part of the evaluation study, these methods are also included.

We propose modifications of Rubin's and Paass's methods, denoted by REG.LOGLIN* and HOD.LOGLIN* respectively, which are based on the log linear method of imputation as introduced by Singh (1988). The proposed modifications use auxiliary information to impose categorical constraints on the matched files obtained from REG* and HOD* methods. In this way, categorical association parameters (estimated via log linear modelling) which measure departure from conditional independence (in the categorical sense) are preserved in the matched file. These categorical constraints are expected to render joint distributions for the completed file A data robust to inferior quality or imperfect nature of the auxiliary data from file C. If auxiliary information is in the form of a categorical distribution and not at the micro-level, then CIA based matching methods can be modified by imposing categorical constraints; in this case the CIA is being used only within X, Y categories. For example, with the usual methods of imputation REG and HOD, which could be used to match by ignoring Y , we can get the corresponding modified versions as REG.LOGLIN and HOD.LOGLIN. These two methods are also considered in this paper.

Note that the categorically constrained matching methods are different from the usual constrained statistical matching methods where the constraints are in the form of a few characteristic measures from file B (such as mean and variance) that variables in the matched file must satisfy. Another key distinction is that the usual constrained matching methods focus on the marginal distribution of Z , whereas the focus here is on the conditional distribution, albeit categorical, which is more relevant for file A; thus there is a basic difference between the two approaches to constrained matching.

Following Rubin (1986) and Paass (1986), we investigate the performance of matching methods empirically. A Monte Carlo study was carried out to investigate the effect of the proposed modifications to the existing methods for the two cases, with and without auxiliary information. This would allow analysis of sensitivity to failure of the CIA and gains from using auxiliary data. The synthetic data for the simulation study was generated from

multivariate normal distributions with some log normal contamination to induce asymmetry. An important advantage of using synthetic data is that relevant control parameters could be modified to yield different distributional scenarios for the matching problem. Eight methods (four existing ones, REG, REG*, HOD, HOD*, and four proposed ones, REG.LOGLIN, REG.LOGLIN*, HOD.LOGLIN, HOD.LOGLIN*) were compared by four evaluation measures (two at the unit level and two at the aggregate level) as mentioned earlier; see Section 6 for details. The main findings of the empirical study can be summarized as follows.

- (i) Use of auxiliary information to avoid the CIA could considerably improve the quality of the matched file. However, if there is no auxiliary information, then among CIA based methods (*i.e.* REG and HOD), the HOD method has better overall performance. Furthermore, an interesting finding was that for small departures from conditional independence, use of auxiliary information may not improve performance of the HOD method with respect to aggregate level evaluation measures. This should have important practical implications in the absence of readily available auxiliary information.
- (ii) The REG* method has very favourable performance with respect to unit level measures. By contrast, it has extremely unfavourable performance with respect to aggregate level measures. This is probably due to the shrinkage towards the mean phenomenon for regressions procedures.
- (iii) The HOD* method does considerably better than REG* at the aggregate level but performs, in general, marginally worse than REG* at the unit level.
- (iv) Categorical constraints, in general, improve performance of REG* and HOD* methods. Specifically, the REG.LOGLIN* method shows slight improvement at the aggregate level, but HOD.LOGLIN* shows considerable improvement at the aggregate level. Their performances at the unit level remain essentially unaffected.
- (v) At the aggregate level, the HOD.LOGLIN method based only on categorical auxiliary information performs generally better than HOD.LOGLIN* based on micro-level auxiliary information. At the unit level, however, HOD.LOGLIN shows marginal deterioration in comparison to HOD.LOGLIN*. This finding may be important from practical considerations because HOD.LOGLIN is computationally much less demanding than HOD.LOGLIN* and does not require micro-level auxiliary information. The REG.LOGLIN method does not have such favourable performance, probably again due to the shrinkage to the mean effect.

(vi) If the auxiliary data is outdated or proxy, there may still be gain in using it. In this context, the HOD.LOGLIN method performs quite favourably and in fact, has fairly robust behaviour with respect to imperfect auxiliary information. Note that since this method uses only information about categorical associations from auxiliary data, it would seem reasonable for this to be affected only slightly by a limited degree of outdatedness or proxyiness in file C. The REG.LOGLIN method, however, does not share this property.

It should be noted that there have been several empirical investigations in the past to evaluate statistical matching methods. Among those that do not consider the use of auxiliary information, some main references are Ruggles, Ruggles and Wolff (1977), Paass and Wauschkuhn (1980), Barr, Stewart and Turner (1981) and Rodgers and DeVol (1982). Paass (1986) provides an excellent review of these empirical tests on the quality of matching methods.

All of the studies cited above confirmed the seriousness of the CIA. This stresses the need for additional information to be incorporated in the matching process. There have been few empirical studies considering the use of auxiliary information and the impact of the CIA; Paass (1986) considered an evaluation with synthetic data only, whereas Armstrong (1989) considered simulations with both synthetic and real data. The present study could be considered as complementary to these studies in the sense that some new methods are included and the choice of underlying population distributions is reasonably broad.

The organization of this paper is as follows. Section 2 describes different types of auxiliary information. A brief review of alternative matching methods using auxiliary information is given in Section 3 and the proposed modifications using categorical constraints are described in Section 4. Different types of matching methods are illustrated in Section 5 by means of a simple numerical example. The description of the design of the empirical study on the proposed matching methods is given in Section 6 and the discussion of results in Section 7. Finally, Section 8 contains concluding remarks and some directions for further research.

2. TYPES OF AUXILIARY INFORMATION

Although a current and sufficiently large micro-datafile with information on the full set of variables is not available, it may be the case that an additional auxiliary source exists containing information on some of the joint relationships of either the full set of variables (X, Y, Z) or perhaps the reduced set (Y, Z) . When this is the case it can be incorporated into the matching process to avoid the CIA and improve the quality of the completed file by reducing distortions in the joint relationships in the matched file.

Such auxiliary information may emanate from various possible sources and may reside in several different forms. Since the purpose of the auxiliary information is only to aid in avoiding the CIA, we limit its use in that information from the host or donor files is never overridden or modified by the auxiliary information. In other words, the objective is to borrow additional information from the auxiliary source not available in the source files. This is accomplished in such a way that confidentiality concerns associated with the auxiliary source would not be violated and implies that the auxiliary source could be a specially conducted small scale survey or a confidential datafile.

Another implication is that the auxiliary information need not be perfect. That is, it may be deficient in some sense. For instance, it may come from an outdated data source (perhaps a previous census or survey), but from which the required auxiliary information may still be valid, or at least represent an improvement over the otherwise default CIA. On the other hand, the auxiliary information may refer to a set of proxy variables expected to behave similarly to the variables of interest.

Auxiliary information could be at the macro-level or micro-level. At the macro-level, it could take the form of either correlations or categorical cell proportions or possibly some other parameters. If the auxiliary information in file C is on the conditional correlation of Y and Z given X , *i.e.* $\rho_{Y,Z|X}$, it can be used with the (X, Y) and (X, Z) correlations from files A and B to estimate the unconditional correlation of (Y, Z) using

$$\rho_{Y,Z} = \rho_{X,Y} \rho_{X,Z} + \rho_{Y,Z|X} (1 - \rho_{X,Y}^2)^{1/2} (1 - \rho_{X,Z}^2)^{1/2}. \quad (2.1)$$

Now data from files A and B can be used to obtain a linear regression of Z on X and Y for the REG* method (see Section 3.1). If auxiliary information on only the unconditional correlation of Y and Z is available, then it can also be used in a similar manner.

The second type of macro-level auxiliary information from file C would be in the form of a categorical distribution for (X^*, Y^*, Z^*) where '*' denotes the categorical transformation of the original variable. If some variables were categorical to begin with, then it may not be necessary to change them. The frequency table required for categorically constrained matching methods can be obtained by raking the (X^*, Y^*, Z^*) table corresponding to file C such that its marginal tables (X^*, Y^*) and (X^*, Z^*) match respectively with the (X^*, Y^*) table from file A and (X^*, Z^*) table from file B. Note that the (X^*, Z^*) table from file B would have to be raked first to match its X^* marginal with that from file A. The method of raking preserves the (Y^*, Z^*) and (X^*, Y^*, Z^*) associations of the (X^*, Y^*, Z^*) table from file C in deriving the categorical constraints. The above adjustment of the (X^*, Y^*, Z^*) table

from file C is reasonable on the grounds that information about the (X^*, Y^*) distribution from file A and about the (X^*, Z^*) distribution from B are believed to be more precise or appropriate than those from file C. If only the (Y^*, Z^*) distribution is available (or used) from file C, then the above raking procedure could be modified to obtain suitable categorical constraints. In this case, the (Y^*, Z^*) association from file C would be preserved and the three factor (X^*, Y^*, Z^*) association term would be assumed to be zero. To achieve this, first the (X^*, Z^*) table from B is raked as before to match the X^* margin from A and then the (Y^*, Z^*) table from C is raked to match the Y^* margin from A and the Z^* margin from B. Then, a three dimensional table of ones is raked to match the (X^*, Y^*) table from A, the adjusted (X^*, Z^*) table from B and the adjusted (Y^*, Z^*) table from C. The categorical counts obtained by these procedures need not be integer values. They are rounded randomly by redistributing fractional counts by sampling cells randomly without replacement with probabilities proportional to the fractions for each cell. This is done independently for each (X^*, Y^*) category.

The next section elaborates on the use of auxiliary information in statistical matching. It also describes the use of auxiliary micro-data. In most cases when micro-level auxiliary information is available, it is possible to roll it up to the macro-level and obtain reliable information on correlations and categorical cell proportions. The validity and reasonableness of this would depend in part on the size of the micro-level datafile.

3. REVIEW OF ALTERNATIVE STATISTICAL MATCHING METHODS

3.1 The Regression Method

We first describe a regression method which uses auxiliary information. This is a version of the method due to Rubin (1986). A parametric form of the regression of Z on X and Y is assumed and the corresponding parameters are then estimated from data in files A, B, and C. For example, in the case of a linear regression, we have the model

$$E(Z | X, Y) = \beta_0 + \beta_1 X + \beta_2 Y, \\ V(Z | X, Y) = \sigma^2, \quad (3.1)$$

where β_0, β_1 , and β_2 are estimated from equations similar to the usual least squares equations by combining information from files A, B, C suitably. Below we describe a procedure for doing this which is somewhat different from the one described in Rubin (1986). If file C has (X, Y, Z) information, then estimates can be obtained of the conditional correlation $\rho_{Y,Z|X}$ from C, the correlation $\rho_{X,Z}$, mean μ_Z , and standard deviation σ_Z from B and the correlation $\rho_{X,Y}$, means μ_X, μ_Y , and standard deviations σ_X, σ_Y from A.

Thus file B will be used only if file A is deficient in information about the quantity of interest and file C will be used for some information only when A and B are deficient. Thus we assume a hierarchy of reliability or relevance of the files A, B, and C. Such a hierarchy was not assumed by Rubin. We can then get the required estimates from

$$\beta_2 = \rho_{Y,Z|X} \frac{\sigma_{Z|X}}{\sigma_{Y|X}}, \quad \beta_1 = \rho_{X,Z|Y} \frac{\sigma_{Z|Y}}{\sigma_{X|Y}}, \\ \beta_0 = \mu_Z - \beta_1 \mu_X - \beta_2 \mu_Y, \quad (3.2)$$

where

$$\sigma_{Z|X} = (1 - \rho_{X,Z}^2)^{1/2} \sigma_Z, \quad \sigma_{Y|X} = (1 - \rho_{X,Y}^2)^{1/2} \sigma_Y, \\ \sigma_{Z|Y} = (1 - \rho_{Y,Z}^2)^{1/2} \sigma_Z, \quad \sigma_{X|Y} = (1 - \rho_{X,Y}^2)^{1/2} \sigma_X, \quad (3.3)$$

and $\rho_{X,Z|Y}$ is obtained from the standard formula after first calculating $\rho_{Y,Z}$ from (2.1), *i.e.*

$$\rho_{X,Z|Y} = (\rho_{X,Z} - \rho_{X,Y} \rho_{Y,Z}) (1 - \rho_{X,Y}^2)^{-1/2}. \quad (3.4)$$

It may be noted that under the normality assumption, departures from conditional independence are parametrized by $\rho_{Y,Z|X}$. Under conditional independence, $\rho_{Y,Z|X} = 0$ and the model (3.1) reduces to the simple linear regression of Z on X , *i.e.*

$$E(Z | X) = \beta_0 + \beta_1 X, \quad V(Z | X) = \sigma^2, \quad (3.5)$$

which can be specified by combining information from files A and B or from file B alone. The formulas (3.2) reduce to

$$\beta_2 = 0, \quad \beta_1 = \rho_{X,Z} \frac{\sigma_Z}{\sigma_X}, \quad \beta_0 = \mu_Z - \beta_1 \mu_X. \quad (3.6)$$

For the case when file C contains information about $\rho_{Y,Z}$ only, the parameters of (3.1) can be easily estimated in a similar manner by combining information from A, B and C.

After the regression model is determined, the REG* method can be applied in the following two steps. Step II is important because we want to have live values of Z so that relationships among components of multivariate Z are preserved.

REG* (Step I) For each (X, Y) in A, find an intermediate value Z_{int} from the regression model (3.1).

REG* (Step II) Replace each (X, Y, Z_{int}) obtained in Step I with (X, Y, Z_{match}) where Z_{match} denotes a live Z -value from B which is nearest under the Euclidean distance in (X, Z) where the components X and Z would be scaled by their respective standard deviations. In other words, the hot deck distance method is used to find the live value. This was termed "regression with predictive mean matching" by Rubin; see Little and Rubin (1987).

Another point of departure from the method described by Rubin (1986) is that in his method a predicted Y is found for records on file B using an equation analogous to (3.1) and then corresponding predicted Z values are found; then records on file A are matched to records on file B based on the difference in predicted Z -values.

If auxiliary information is not available then the matching method REG under CIA can be used. The two steps are

REG (Step I) For each (X, Y) in A, find Z_{int} from the simple regression model (3.5).

REG (Step II) Same as in REG*.

The method described by Rubin (1986) differs in that a predicted Z is also obtained for records on file B using (3.5), and then records on file A are matched to records on file B based on the difference in predicted Z -values. In the present example, where X is univariate, this is equivalent to matching on X .

3.2 The Hot Deck Method

We first describe a hot deck method using auxiliary data. This is a version of the method due to Paass (1986). Here, ideas of nonparametric regression are used. In parametric regression, the conditional distribution of Z given X and Y is specified in a wide sense by mean and variance functions in terms of a few parameters. In nonparametric regression the techniques of nonparametric density estimation are used to estimate the conditional distribution itself. For instance, in the case of the nearest neighbour method of density estimation, for each (X, Y) , K nearest neighbours (with respect to a distance function such as the Euclidean distance in (X, Y) are determined and then the conditional distribution is represented by this sample (possibly weighted) of the K neighbours where K is an integer specified suitably. Thus, $P(Z \in U | X, Y)$ can be specified as a conditional expectation,

$$E(I_U(Z) | X, Y) = \sum_{i=1}^K w_i(X, Y) I_U(Z_i), \quad (3.7)$$

where w_i 's denote weights which decrease with growing distance of (X_i, Y_i) from (X, Y) and I_U is the indicator function for the set U .

In Paass's method, first the conditional distribution of Z for each (X, Y) in A is determined by representing it with a set of K Z -values using nonparametric regression. In other words, K Z -values are added to each (X, Y) . Then for each (X, Y) in A, a single live Z -value, Z_{match} , from B is obtained which is nearest under (X, Z) -distance. This gives the matched file with (X, Y, Z_{match}) . The conditional distributions for file A are obtained by an iterative process in the case of file C with (Y, Z) information, as follows. Choose K initial values for nearest neighbours for Z in file A,

for Y in file B, and for X in file C. This can be done by the usual hot deck method of imputation. Now each cycle consists of determining conditional distributions for elements (X, Y) in A from information in C, *i.e.* suitably updating K Z -values in A from Z -values in C using (X, Y) distance, and then updating K Y -values in file B from those of file A using (X, Z) distance, and finally updating K X -values in C from those of file B using (Y, Z) distance. This cycle is repeated until the maximal difference between some statistics for the three-dimensional distribution of (X, Y, Z) of successive iterations (*e.g.* covariance matrix) falls below a given threshold. At convergence, each file has K added values representing respective conditional distributions. In the other case in which file C has information about (X, Y, Z) the process becomes noniterative. We simply use file C to get K Z -values for A using (X, Y) distance and then get Z_{match} from B for each (X, Y) in A using (X, Z) -distance. This case was, however, not considered by Paass.

In the empirical study considered in this paper we did not use the above iterative version of Paass's method when file C had (Y, Z) data, because of its computationally intensive nature. Instead, we used a simplified noniterative version with $K = 1$. This method, denoted by HOD*, consists of the following two steps.

HOD* (Step I) For each (X, Y) in A, find an intermediate value Z_{int} from C using hot deck with Y -distance in the case of (Y, Z) auxiliary information and with (X, Y) Euclidean distance in the case of (X, Y, Z) auxiliary information.

HOD* (Step II) Replace each (X, Y, Z_{int}) obtained in Step I by (X, Y, Z_{match}) where Z_{match} is obtained from B using hot deck with (X, Z) Euclidean distance.

If file C were not available, then the matching method HOD under CIA can be used. The two steps for HOD are

HOD (Step I) Determine suitable X -categories as in usual hot deck imputation.

HOD (Step II) For each (X, Y) in A, impute a live Z -value from the corresponding X -category from B using hot deck with X -distance.

4. THE PROPOSED MODIFICATIONS BY CATEGORICALLY CONSTRAINED MATCHING

We propose modifications to REG, REG*, HOD and HOD* matching methods by imposing categorical constraints on the Z -values selected from B for completing A. The purpose of these constraints is to preserve categorical associations (as defined by log linear modelling) under a suitable partition of (X, Y, Z) for the matched file. These

associations are obtained by combining information from A, B and C. The idea of categorically constrained matching is based on the method of log linear imputation (*cf.* Singh 1988, Singh *et al.* 1988). Here the constraints could be based on auxiliary information which could be used to estimate the categorical conditional distribution, or some aspects of it, but which would not be of sufficient quality to estimate the full conditional distribution.

We start with a suitable partition of X, Y and Z variables. Let X^*, Y^*, Z^* denote the corresponding categorically transformed variables. Now the distribution of cell proportions for the (X^*, Y^*, Z^*) table can be parametrized by a log linear model

$$\log p_{ijk} = u + u_{1i} + u_{2j} + u_{3k} + u_{12ij} + u_{13ik} + u_{23jk} + u_{123ijk}, \quad (4.1)$$

where p_{ijk} denotes the proportion for (i, j, k) th cell and 1, 2, 3 denote respectively X^*, Y^* , and Z^* . It should be noted that the parametrization (4.1) holds for arbitrary underlying distributions of the original variables (X, Y, Z) . The files A and B, of course, do not contain any information about the two-factor effects u_{23} and three-factor effects u_{123} . If these are set to zero, this amounts to assuming CIA in the categorical sense, *i.e.* $Y^* \perp Z^* \mid X^*$. However, with auxiliary information in file C, this assumption can be avoided because the parameters u_{23} and u_{123} could be estimable from C. Thus, regardless of the form of the joint distribution of (X, Y, Z) , the above log linear modelling provides a unified approach for gauging departures from CIA at least in the categorical sense. In the linear regression approach, on the other hand, departures from CIA are parametrized by $\rho_{Y,Z|X}$ only in the case of normality.

As was explained in Section 2, the auxiliary information from file C (either on (Y, Z) or on (X, Y, Z)) is first used to construct categorical constraints in the form of a (X^*, Y^*, Z^*) distribution. This is done by means of raking such that u_{23} and u_{123} effects from file C are preserved. The categorically constrained version of REG*, denoted by REG.LOGLIN*, can now be defined by the following two steps.

REG.LOGLIN* (Step I) Same as in Step I of REG*.

REG.LOGLIN* (Step II) Same as in Step II of REG* except that categorical constraints are imposed, implying that match order is required when obtaining live Z -values from B. We first find the match with minimum distance in (X, Z) . The (X^*, Y^*, Z^*) category of the completed record would be noted and if the resulting number of matched records in that (X^*, Y^*, Z^*) category does not exceed the count imposed by the categorical constraints that match is allowed. Otherwise, that match is rejected and the match with the second smallest distance is examined.

The process continues until file A is completed, and then the distribution of (X^*, Y^*, Z^*) in the completed file must satisfy the categorical constraints.

Similarly, the categorically constrained version of HOD*, denoted by HOD.LOGLIN*, consists of the following two steps.

HOD.LOGLIN* (Step I) For each (X, Y) in A, find an intermediate value, Z_{int} , from C using hot deck with Y -or (X, Y) -distance as the case may be such that the categorical constraints are satisfied. This step is similar to Step II of REG.LOGLIN*.

HOD.LOGLIN* (Step II) For each (X, Y, Z_{int}) , a live value, Z_{match} , from B is determined using hot deck with (X, Z) -distance while respecting the category of Z_{int} .

An alternative approach for HOD.LOGLIN* would have been to impute an intermediate Z_{int} without constraints and then to use categorically constrained distance matching to get a live value from file B, as in Step II of REG.LOGLIN*. This was also tried but did not work well so it was dropped from the study because of computational burden. One possible explanation for its poor performance is shrinkage to the mean for the Z_{int} values from file C due to file C being too small. That is, the Z_{int} values would tend to be near the centre of the distribution and when the categorical constraints are then imposed the final Z values would tend to be clumped at the inside boundaries of the outer Z categories.

Suppose file C has information only at the macro-level in the form of a categorical distribution, or the micro-level information in C is considered unreliable but the information in the categorical distribution under a somewhat coarse partition is considered reliable. We can then define categorically constrained versions of the REG and HOD methods, to be denoted by REG.LOGLIN and HOD.LOGLIN respectively. The two steps for REG.LOGLIN are

REG.LOGLIN (Step I) Same as in Step I of REG.

REG.LOGLIN (Step II) Same as in Step II of REG.LOGLIN*.

Similarly, HOD.LOGLIN consists of the following two steps.

HOD.LOGLIN (Step I) Same as in Step I of HOD.

HOD.LOGLIN (Step II) Same as in Step II of REG.LOGLIN* except that no intermediate values Z_{int} exist, so that matching is based on X -distance instead of (X, Z) -distance.

For both REG.LOGLIN and HOD.LOGLIN, which do not require micro-level information on file C, the CIA is being used only within X, Y categories. Thus a reduced form of conditional independence is being assumed and the consequences of this assumption should not be as severe as those of the full CIA.

5. AN ILLUSTRATIVE EXAMPLE

Before we investigate the empirical properties of the proposed modifications in relation to the previously proposed methods, it may be instructive to consider a simple numerical example to illustrate the types of computation involved with the eight methods. Suppose files A, B and C are as shown in Table 1 which are based on random samples drawn from a multivariate normal with mean 0 and covariance matrix specified by $\sigma_X = \sigma_Y = \sigma_Z = 1$, $\rho_{X,Y} = \rho_{X,Z} = .5$ and $\rho_{Y,Z} = .7$ (which implies that $\rho_{Y,Z|X} = .6$). Here, file C is assumed to have only (Y,Z) information. For file A, Z-values are suppressed in Table 1 but are shown in Table 3 for computing evaluation measures. Suppose we employ, for simplicity and in view of small file sizes, a rather coarse categorical transformation for X,Y,Z by considering only two categories, $(-\infty, 0)$ and $[0, \infty)$. Then, the three two dimensional count tables corresponding to files A, B and C can be constructed as in Table 2(a). Table 2(b) shows the adjusted tables for B and C so that they match the appropriate marginals as described in Section 2. Table 2(c) gives the three-dimensional

table obtained after raking and Table 2(d) gives the desired categorical constraints after random rounding of entries of Table 2(c) as explained earlier in Section 2.

The eight methods were applied to the data of Table 1 and the matching results are shown in Table 3 along with the true values of Z which were suppressed in Table 1.

The evaluation measures shown in Table 3 were briefly introduced earlier in the introduction and are fully explained in the next section. The categorical partition for the χ^2 measure was the same as the one used for deriving categorical constraints. Note that since the partitioning is not changed for evaluation, the χ^2 values for M3, M4, M7 and M8 would be identical. It should be pointed out that the evaluation measures are given only for the sake of illustrating the calculation and should not be construed as indicators for the relative performance of various methods because they are based on just one small sample realization.

The method M8 (HOD.LOGLIN*) happens to be the most computationally intensive, the details of which are shown in Table 4. From this, it would be relatively easy to visualize the computational steps required for other methods.

Table 1
Data for Files A, B, C

Record Identifier	File A		Record Identifier	File B		Record Identifier	File C	
	X	Y		X	Y		Y	Z
A1	-0.86	-0.32	B1	-0.95	-0.69	C1	-0.40	-0.60
A2	-0.77	-0.33	B2	-0.64	-0.83	C2	-2.33	-2.81
A3	-0.09	-0.26	B3	-1.58	-0.11	C3	-0.79	-0.47
A4	-0.42	0.62	B4	-0.42	0.36	C4	0.67	-0.29
A5	-0.81	0.56	B5	0.97	-0.42	C5	-0.65	1.19
A6	-0.56	0.00	B6	1.09	-1.16	C6	-1.32	0.05
A7	0.37	-0.04	B7	0.44	-0.49	C7	-0.55	0.70
A8	0.06	-1.29	B8	0.14	-0.38	C8	0.55	0.66
A9	0.95	-2.15	B9	1.33	1.24	C9	1.31	1.12
A10	1.90	-1.07	B10	0.80	0.85	C10	1.46	2.58
A11	1.32	0.61	B11	1.60	0.31			
A12	1.38	0.79	B12	1.42	0.99			
A13	1.63	1.03						
A14	0.50	1.24						
A15	0.90	1.19						

Table 2
Categorical distributions for files A, B, C under the given $2 \times 2 \times 2$ partition

(a)	File A		File B			File C		
	$Y < 0$	$Y \geq 0$	$Z < 0$	$Z \geq 0$	$Z < 0$	$Z \geq 0$		
$X < 0$	3	3	$X < 0$	3	1	$Y < 0$	3	3
$X \geq 0$	4	5	$X \geq 0$	4	4	$Y \geq 0$	1	3

(b)	Unadjusted File A Table		Adjusted File B Table			Adjusted File C Table		
	$Y < 0$	$Y \geq 0$	$Z < 0$	$Z \geq 0$		$Z < 0$	$Z \geq 0$	
$X < 0$	3	3	$X < 0$	4.5	1.5	$Y < 0$	5.15	1.85
$X \geq 0$	4	5	$X \geq 0$	4.5	4.5	$Y \geq 0$	3.85	4.15

(c)	Raked $2 \times 2 \times 2$ table of ones to match the marginals in Table 2(b)			
	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2.55	1.95	0.45	1.05
$X \geq 0$	2.60	1.90	1.40	3.10

(d)	Categorical constraints by randomly rounding entries of Table 2(c)			
	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2	2	1	1
$X \geq 0$	2	2	2	3

Table 3
Comparison of Eight Matching Methods for Completing File A

File A			Matched Z-Values							
			Versions of REG Method				Versions of HOD method			
<i>X</i>	<i>Y</i>	<i>Z</i>	M1	M2	M3	M4	M5	M6	M7	M8
−0.86	−0.32	−0.97	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69
−0.77	−0.33	0.16	−0.69	−0.69	−0.69	−0.69	−0.83	−0.69	−0.83	−0.69
−0.09	−0.26	0.19	−0.38	−0.38	0.36	0.36	0.36	−0.38	0.36	0.36
−0.42	0.62	−0.44	−0.38	0.36	0.36	−0.38	0.36	−0.38	0.36	−0.38
−0.81	0.56	−0.76	−0.69	0.36	−0.69	0.36	−0.69	0.36	−0.69	0.36
−0.56	0.00	1.06	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83
0.37	−0.04	−1.18	−0.38	−0.38	−0.38	0.36	−0.49	−0.49	−0.49	−0.49
0.06	−1.29	0.33	−0.38	−0.38	−0.36	−0.38	−0.38	−0.38	0.85	0.36
0.95	−2.15	−1.26	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16
1.90	−1.07	0.01	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
1.32	0.61	2.08	0.31	0.99	0.31	0.99	1.24	−0.42	1.24	−0.42
1.38	0.79	0.32	0.31	0.99	0.31	0.99	0.99	−0.42	0.99	−0.42
1.63	1.03	1.53	0.31	0.99	0.31	0.99	0.31	0.99	0.31	0.99
0.50	1.24	1.34	−0.49	0.85	−0.49	−0.38	−0.49	0.85	−0.49	0.85
0.90	1.19	−1.01	−0.42	0.85	−0.42	−0.42	−0.42	0.85	−0.42	0.85
Evaluation Measures		MAD-Z	0.79	0.81	0.76	0.78	0.79	0.85	0.78	0.78
		χ^2	13.07	13.34	1.75	1.75	2.70	10.78	1.75	1.75

Note: M1: REG M2: REG* M3: REG.LOGLIN M4: REG.LOGLIN* M5: HOD M6: HOD* M7: HOD.LOGLIN M8: HOD.LOGLIN*.

Table 4
Computational Steps Required for M8 (HOD.LOGLIN*)

(<i>X</i> , <i>Y</i>) cell	<i>X</i>	<i>Y</i>	Match Order	<i>Y</i> - dist	<i>Z</i> _{int}	(<i>X</i> , <i>Z</i>)- dist	<i>Z</i> _{match}
<i>X</i> < 0	−0.86	−0.32 (A1)	2	.08	−0.60 (C1)	.12	−0.69 (B1)
<i>Y</i> < 0	−0.77	−0.33 (A2)	1	.07	−0.60 (C1)	.19	−0.69 (B1)
	−0.09	−0.26 (A3)	3	.29	0.70 (C7)	.47	0.36 (B4)
<i>X</i> < 0	−0.42	0.62 (A4)	2	.05	−0.29 (C4)	.56	−0.38 (B8)
<i>Y</i> ≥ 0	−0.81	0.56 (A5)	1	.01	0.66 (C8)	.50	0.36 (B4)
	−0.56	0.00 (A6)	3	.40	−0.60 (C1)	.25	−0.83 (B2)
<i>X</i> ≥ 0	0.37	−0.04 (A7)	4	.36	−0.60 (C1)	.14	−0.49 (B7)
<i>Y</i> < 0	0.06	−1.29 (A8)	1	.03	0.05 (C6)	.57	0.36 (B4)
	0.95	−2.15 (A9)	2	.18	−2.81 (C2)	1.66	−1.16 (B6)
	1.90	−1.07 (A10)	3	.25	0.05 (C6)	.40	0.31 (B11)
<i>X</i> ≥ 0	1.32	0.61 (A11)	1	.06	−0.29 (C4)	.37	−0.42 (B5)
<i>Y</i> ≥ 0	1.38	0.79 (A12)	3	.12	−0.29 (C4)	.43	−0.42 (B5)
	1.63	1.03 (A13)	5	.28	1.12 (C9)	.25	0.99 (B12)
	0.50	1.24 (A14)	2	.07	1.12 (C9)	.41	0.85 (B10)
	0.90	1.19 (A15)	4	.12	1.12 (C9)	.29	0.85 (B10)

Note: Match order between records in files A and C is within (*X*, *Y*) cell under the categorical constraints given by Table 2(d).

6. EMPIRICAL INVESTIGATION OF PROPOSED MATCHING METHODS

This section presents the details of an empirical evaluation through an extensive simulation study with synthetic data generated from symmetric as well as skewed multivariate distributions. Symmetry was introduced via normal distributions, while skewness was introduced via contaminations by multivariate log normal distributions. The reason for using synthetic data is to have control over all of the relevant parameters, including those specifying the joint relationships of the different variables. This permits evaluation of the various approaches to matching as the joint relationships are allowed to depart in a systematic manner from conditional independence. It also permits comparisons of the methods as the underlying distribution generating the data moves away from symmetry. Proxy auxiliary information was generated by changing parameters of the normal distribution generating file C or by inducing log normal contaminations. We thus have four types of matching problems; the two corresponding to symmetric and skewed distributions with nonproxy data for C and the two corresponding to symmetric distributions with two types of proxy data for C. Programming was done on micro-computers using the software GAUSS.

6.1 Design of the Monte Carlo Study

In order to simulate statistical matching three datafiles are needed: a host file A, a donor file B, and an auxiliary file C. These are generated synthetically from specified distributions, with each file containing the three variables X , Y and Z . In file A the variable Z is suppressed and in file B the variable Y is dropped. The suppressed Z -values in file A are used to evaluate the performance of the various methods of statistical matching. File C could have only (Y, Z) information (by suppressing X) or the full (X, Y, Z) information. The empirical results presented in this paper correspond to file C with only (Y, Z) variables although file C with (X, Y, Z) variables was also included in the study (see Singh *et al.* 1990).

Runs of 100 simulations apiece were performed for each combination of design parameters considered. Four evaluation measures were calculated for each simulation and then were combined over all 100 simulations.

Files A and B were always generated from the same underlying distribution, with each containing 500 independent and identically distributed observations. File C contained 250 observations, not necessarily from the same distribution as that for files A and B; that is, file C could contain either proxy or nonproxy auxiliary information.

The distribution of observations (X, Y, Z) was multivariate normal with some log normal contamination introduced by taking the exponentials of X , Y and Z for

some of the observations. Individual observations were contaminations or not according to a Bernoulli process with probability fixed for any particular run of 100 simulations. Prior to contamination X , Y and Z were standard normal. The covariances of (X, Y) and (X, Z) prior to contamination were always .5, with the covariance of (Y, Z) varying from run to run. Consequently, the conditional correlation of Y and Z given X , $\rho_{Y,Z|X}$, was also varied from run to run.

For most runs the distribution of observations in the auxiliary file C was the same as that in files A and B. However, if in an application the source of auxiliary information is historical or via proxy variables this assumption may be unreasonable. Two series of runs were carried out with proxy auxiliary information. In the first series the auxiliary data had a different $\rho_{Y,Z|X}$. In the second series the auxiliary data had some log normal contamination.

For the proposed methods which use categorical constraints and for defining matching categories for the HOD method, it was necessary to choose a categorical partition. Two partitions were used. The first, called standard interval, divided the ranges of the X , Y and Z variables into the categories < -1 , $[-1, 0)$, $[0, 1)$, ≥ 1 ; that is, the partition was centred on the mean of the marginal distribution before contamination, with break points at the centre and at plus or minus one standard deviation. The second partition, called equal probability, was similar but had break points at the quartiles of the pre-contamination marginal distributions; that is, the partition had the categories $< -.6745$, $[-.6745, 0)$, $[0, .6745)$, $\geq .6745$. The partitions were defined in terms of the pre-contamination distributions; for simplicity the same partitions were used when there were log normal contaminations. It would, however, have been more realistic to let the partitions be data dependent.

6.2 The Matching Methods

The eight methods as defined earlier were considered. Except for REG and HOD, all others use auxiliary information. Thus, we have two variants for each depending on whether (Y, Z) or (X, Y, Z) information is available in file C. For the methods HOD and HOD.LOGLIN, three versions of hot deck (namely, rank, random, and X -distance) were considered for finding live Z -values from B although only results based on X -distance are reported here. For the other six methods, although we considered three types of hot deck (namely, Z -distance, (X, Z) -distance, and (X, Y, Z) -distance) for finding live Z -values from B, we show only results for (X, Z) distance here for simplicity. Section 7.3 does contain a brief description of results obtained with different distance measures. The report by Singh *et al.* (1990) contains other details not included here. It may be noted that for using hot deck with (X, Y, Z) -distance to get a live Z -value from B, intermediate Y -values

would have to be first obtained for B from file C, analogous to Z_{int} for file A. Note also that the Euclidean distance was always employed whenever hot deck with distance metric was used. However, variables were not preadjusted by their standard deviations for convenience and because all the variables in the synthetic population had common variances.

6.3 The Evaluation Measures

Four evaluation measures were used to measure how well the different matching methods performed. All of the evaluations are based on comparisons of the matched file to the file with the suppressed true Z -values. Two of the measures are based on categorical comparisons, but the categories used for evaluations need not be the same as those used for categorical constraints by the LOGLIN procedures. The results reported here correspond to using the equal probability partition (see Section 6.1) for matching and the standard interval partition for evaluations. The first of the four evaluation measures is based on unit by unit comparison of the matched and suppressed Z -values. However, the objective of a statistical matching procedures cannot be to reproduce the suppressed Z -values exactly, but to produce Z -values that come from the same distribution given what is known, in this case given X and Y . The last three evaluation measures are based more on comparisons of the conditional distributional properties of Z .

(i) Average of Mean Absolute Differences of Z ($\overline{\text{MAD-Z}}$)

The simplest measure of performance is the mean absolute difference between the matched and suppressed Z -values for records in file A. Monte Carlo averages of these means as well as standard errors were obtained.

The formula for the MAD- Z statistic for any given simulation, is

$$\text{MAD-Z} = \sum_i |Z_{s,i} - Z_{m,i}| / 500, \quad (6.1)$$

where $Z_{s,i}$ is the suppressed Z -value for the i th record in file A, $Z_{m,i}$ is the matched Z -value, and the sum is over all 500 records of file A. MAD- Z denotes the average of the MAD- Z statistics over simulations.

(ii) Average of Absolute Difference of Covariances (AD-Cov)

The second measure of performance is the absolute difference of the conditional covariances of Y and Z given X in the matched and suppressed files. Monte Carlo averages of these absolute differences as well as standard errors were obtained.

For a file with variables X , Y and Z we may define

$$\text{Cov}(Y, Z | X) = \text{Cov}(Y, Z) -$$

$$\text{Cov}(X, Y)\text{Cov}(X, Z)/\text{Var}(X), \quad (6.2)$$

where Cov and Var are the sample covariance and variance operators respectively. In the multivariate normal case this corresponds to the covariance of Y and Z given X . Otherwise it may be interpreted as the covariance of the residuals of a linear regression of Y on X with the residuals of a linear regression of Z on X . The AD-Cov statistic for any given simulation, would be the absolute difference between these quantities for the matched and suppressed files. AD-Cov denotes as usual the average over simulations.

(iii) Average of Chi-square Statistics ($\overline{\chi^2}$)

The third measure of performance, based on categorical comparisons, is a distance measure based on the Pearson chi-square statistic. What is reported is the average chi-square statistic over the 100 simulations, transformed to lie in the interval (0,1).

The formula for the chi-square statistic, is

$$\chi^2 = \sum_{i,j,k} (m_{ijk} - n_{ijk})^2 / (m_{ijk} + .5), \quad (6.3)$$

where m_{ijk} is the number of records in X^* category i , Y^* category j , and Z^* category k in the matched file, n_{ijk} is the same for the suppressed file, and the sum is over all (X^*, Y^*, Z^*) categories. A constant .5 is added to all of the denominators in this sum to avoid the problem of zeros.

Once the mean of the chi-square statistics from 100 simulations, say $\overline{\chi^2}$, is obtained, it is transformed to lie in the interval (0,1) using the transformation (see Bishop, Fienberg and Holland 1975, p 383; here 500 is the size of file A)

$$\text{Transformed } \chi^2 = \{ \overline{\chi^2} / (\overline{\chi^2} + 500) \}^{1/2}. \quad (6.4)$$

(iv) Likelihood Ratio Test (LRT)

The final measure of performance is also based on categorical comparisons. Within each (X^*, Y^*) category that has a minimum number of observations (in the present study, we set it at 20) a likelihood ratio test that the categorical Z -values from the matched and suppressed files come from the same multinomial distribution is performed. The tests for different (X^*, Y^*) categories are then combined to obtain an overall P -value. What is reported is the proportion of times, out of 100 simulations, that the overall P -value was less than .05. The larger this proportion, the greater the difference between the true and matched categorical distributions of Z^* given the (X^*, Y^*) categories.

The minimum sample size of 20 for (X^*, Y^*) categories in file A was required so that the chi-square approximation to the distribution of the test statistic might be reasonable. If the number of Z^* categories was increased, this minimum sample size might also need to be increased.

Using the same notation as in the previous measure, the formula for the likelihood ratio test statistic from the (i, j) (X^*, Y^*) category is

$$\begin{aligned} \text{LRT} = 2 \sum_k \{ & (n_{ijk} + .5) \ln((n_{ijk} + .5) / \\ & (n_{ijk} + m_{ijk} + 1)) + (m_{ijk} + .5) \\ & \ln((m_{ijk} + .5) / (n_{ijk} + m_{ijk} + 1)) \} \\ & + (4n_{ij} + 2K) \ln 2, \end{aligned} \quad (6.5)$$

where

$$\begin{aligned} n_{ij} = \sum_k n_{ijk} = \sum_k m_{ijk}, \quad i = 1, \dots, I, \\ j = 1, \dots, J, \quad k = 1, \dots, K. \end{aligned} \quad (6.6)$$

The asymptotic distribution of this statistic, when the m_{ijk} 's and n_{ijk} 's come from the same multinomial distribution, is chi-square with $(K - 1)$ degrees of freedom. An overall P -value is obtained by adding these statistics and their degrees of freedom for each (X^*, Y^*) category meeting the minimum sample size criterion, and finding the probability of a chi-square variable with the appropriate degrees of freedom being larger than the observed value.

7. RESULTS OF THE MONTE CARLO STUDY

In this section we describe the results of the simulation study. A more complete description is given in Singh *et al.* (1990). Tables of actual numbers underlying Figures 1 through 5 are available upon request.

We have not paid much attention to Monte Carlo standard errors of the evaluation measures in the presentation. This is because they were generally quite small, for example, coefficients of variation were generally less than two percent for the $\overline{\text{AD-Cov}}$ evaluation measure. Furthermore, the evaluations of different methods would be expected to be positively correlated so that the relative differences between matching methods would be even more precisely estimated than suggested by the standard errors. A further indication of the quality of the Monte Carlo evaluations of the various methods is the general smoothness of observed trends, for example, see Figures 2 to 5. In short, any discernible difference in the figures is likely to indicate a real difference.

7.1 Methods with no Auxiliary Information (REG and HOD)

Figures 2 through 5 show how departures from conditional independence affect performance of matching methods which use CIA. Apparently the use of such methods may result in serious bias in the joint relationship of (X, Y, Z) in the matched file. For example, Figure 2 shows a progressive deterioration as the true conditional correlation, $\rho_{Y,Z|X}$, moves away from zero with respect to all measures except $\overline{\text{MAD-Z}}$ which actually shows no deterioration at all. It may be due to the fact that $\overline{\text{MAD-Z}}$ is an unconditional measure which is based on unit by unit comparison of the matched and suppressed Z -values, while the other measures are based on comparisons of the conditional distributions of Z . It is interesting to note from Figure 2 that when the true value of $\rho_{Y,Z|X}$ is small, the performance of the HOD* method, which uses auxiliary information, can be worse with respect to the categorical or aggregate level evaluation measures than the performance of the HOD method which does not make use of auxiliary information. The point at which the use of auxiliary information would become advantageous would depend on the precision of the auxiliary information.

7.2 Methods with Auxiliary Information

Our empirical results do confirm, as expected, that the use of auxiliary information does protect against the failure of the CIA. The degree of protection would depend on the method and the type of auxiliary information used. A brief summary of performances of various methods was presented earlier in the introduction. Here, we will provide some details based on Figures 2 to 5.

In the regression family, the methods using auxiliary information on conditional correlations, namely REG* and REG.LOGLIN*, show very favourable performance with respect to the unit level measures (*i.e.* $\overline{\text{MAD-Z}}$ and $\overline{\text{AD-Cov}}$) for symmetric populations (see Figure 2). They continue to outperform hot deck methods for skewed populations (Figure 3) although the bias tends to increase as the degree of skewness grows. However, for proxy auxiliary information having different conditional correlation (Figure 4), the regression methods perform in a mixed fashion, *i.e.* they could be better or worse than hot deck methods at the unit level. In fact, for the second type of proxy auxiliary information (namely, with log normal contamination; see Figure 5), they tend to be slightly inferior to the HOD.LOGLIN method with respect to the $\overline{\text{AD-Cov}}$ measure. If we restrict ourselves to the regression family, then the REG* method can be recommended with regard to the unit level evaluation measures. However, with respect to the aggregate level, all regression methods show very unfavourable performance. This can probably be explained by the shrinkage to the mean effect as discussed in subsection 7.3.

Matching Methods for Figures 1 to 5

REG	Z_{int} obtained from regression of Z on X , Z_{match} based on (X, Z) distance
REG*	Z_{int} obtained from regression of Z on X and Y , Z_{match} based on (X, Z) distance
REG.LOGLIN	Z_{int} obtained from regression of Z on X , Z_{match} based on (X, Z) distance using categorical constraints
REG.LOGLIN*	Z_{int} obtained from regression of Z on X and Y , Z_{match} based on (X, Z) distance using categorical constraints
HOD	Hot deck using X distance within X categories
HOD*	Z_{int} obtained from file C using hot deck with Y distance, Z_{match} obtained from file B using hot deck with (X, Z) distance
HOD.LOGLIN	Hot deck using X distance within X categories and using categorical constraints
HOD.LOGLIN*	Z_{int} obtained using hot deck with Y distance and using categorical constraints, Z_{match} obtained using hot deck with (X, Z) distance within (X, Y, Z) categories

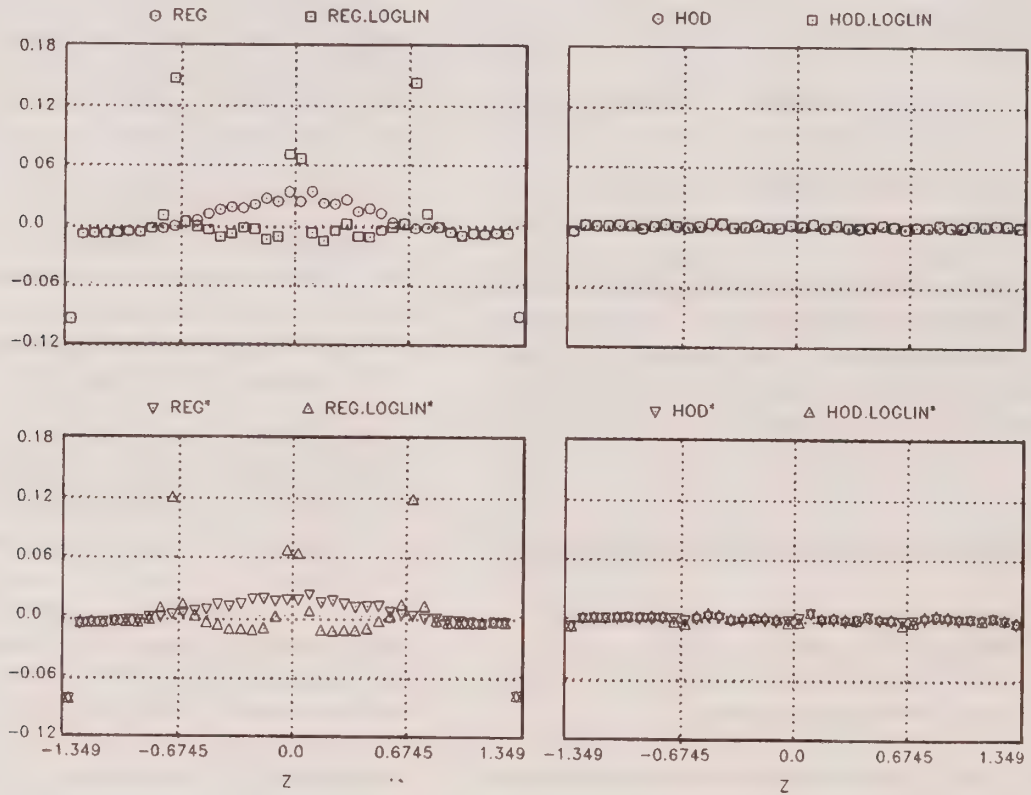


Figure 1. Difference of matched and suppressed marginal Z -histograms (symmetric data, $\rho_{Y,Z|X} = .4$)

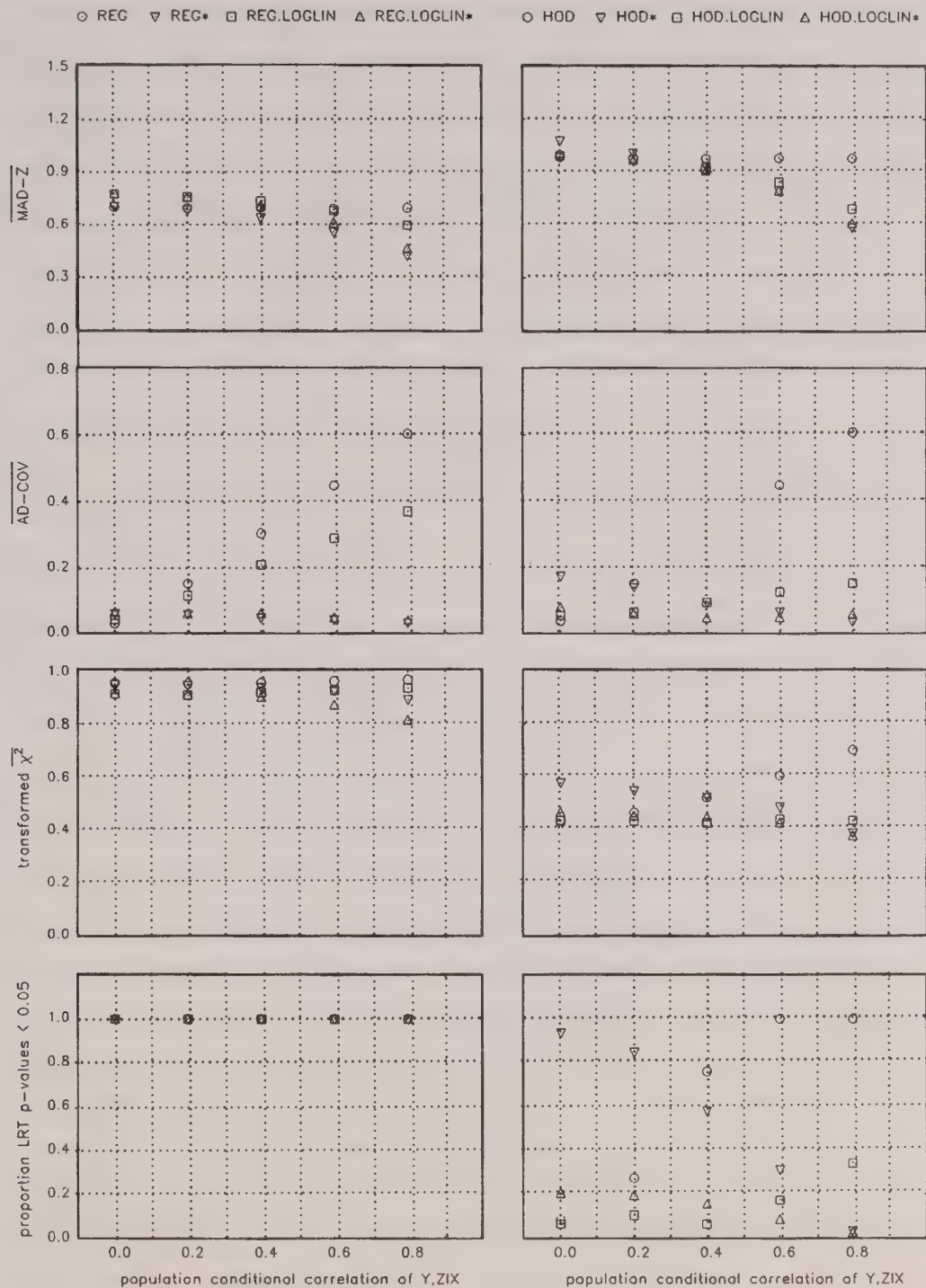


Figure 2. Comparison of statistical matching methods as $\rho_{Y,Z|X}$ varies for the symmetric population, non-proxy auxiliary information

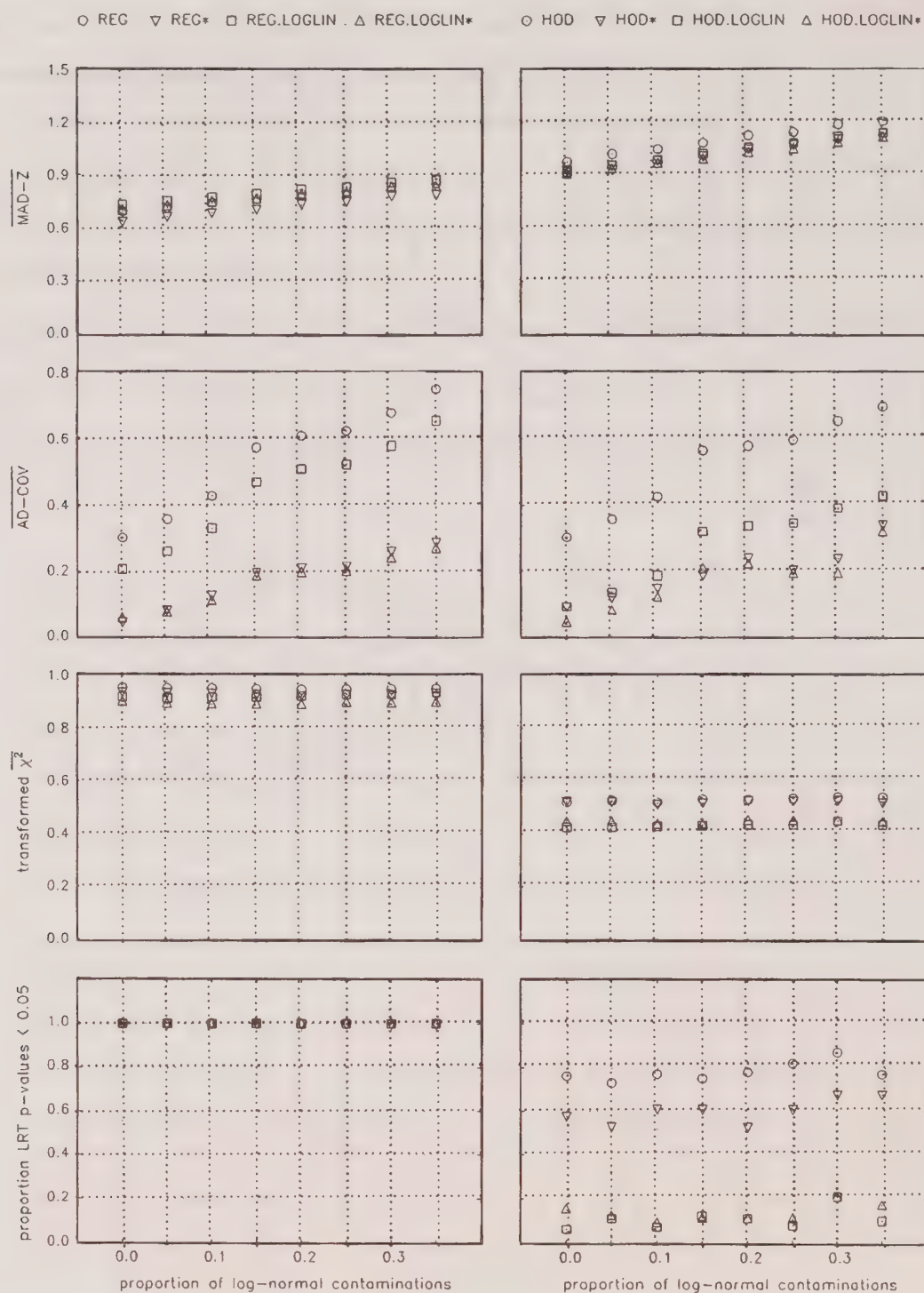


Figure 3. Comparison of statistical matching methods as the proportion of log-normal contamination varies ($\rho_{Y,Z|X}$ before contamination), non-proxy auxiliary information

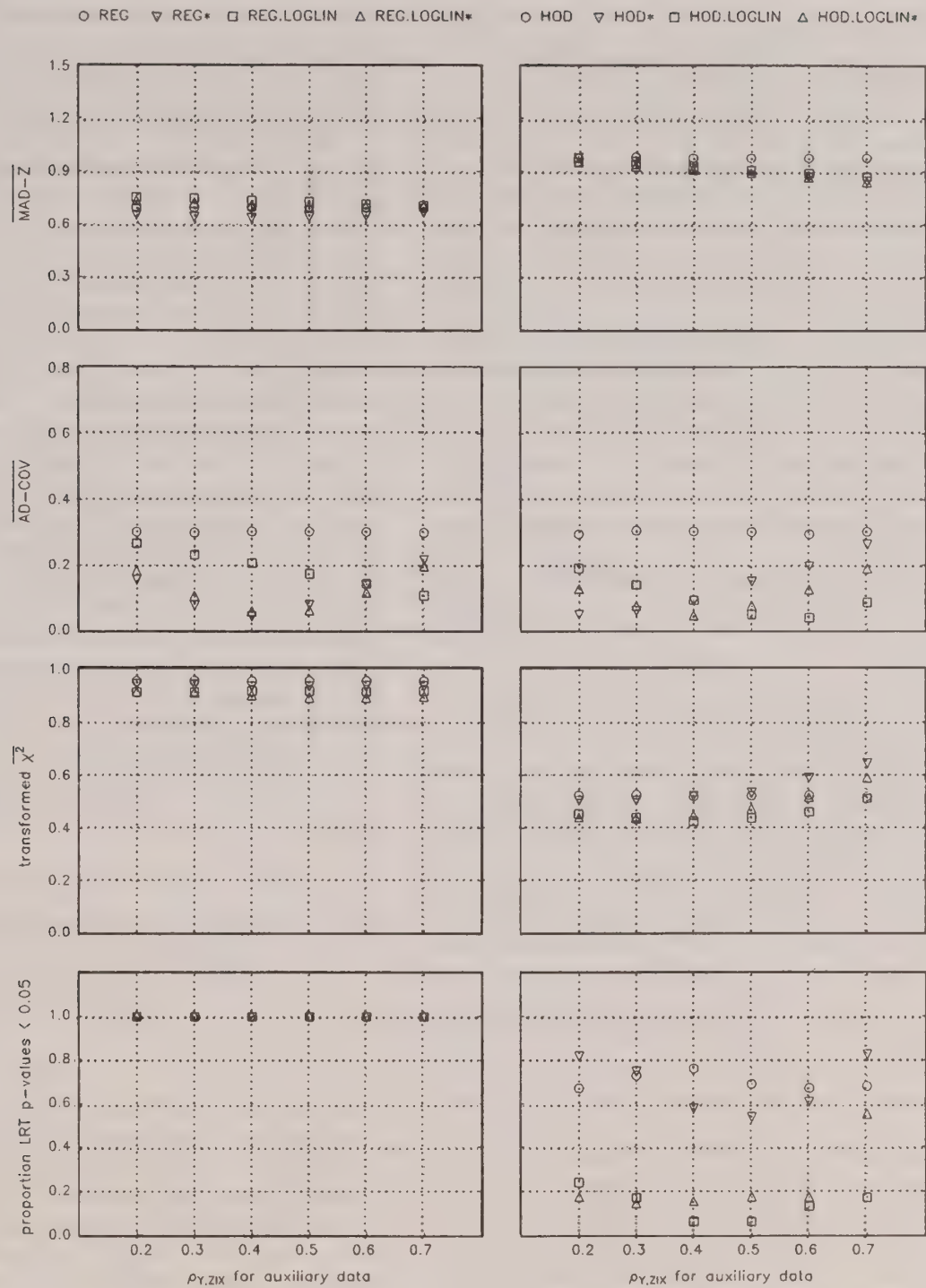


Figure 4. Comparison of statistical matching methods as $\rho_{Y,Z|X}$ varies for the auxiliary data file C ($\rho_{Y,Z|X} = .4$ for files A and B)

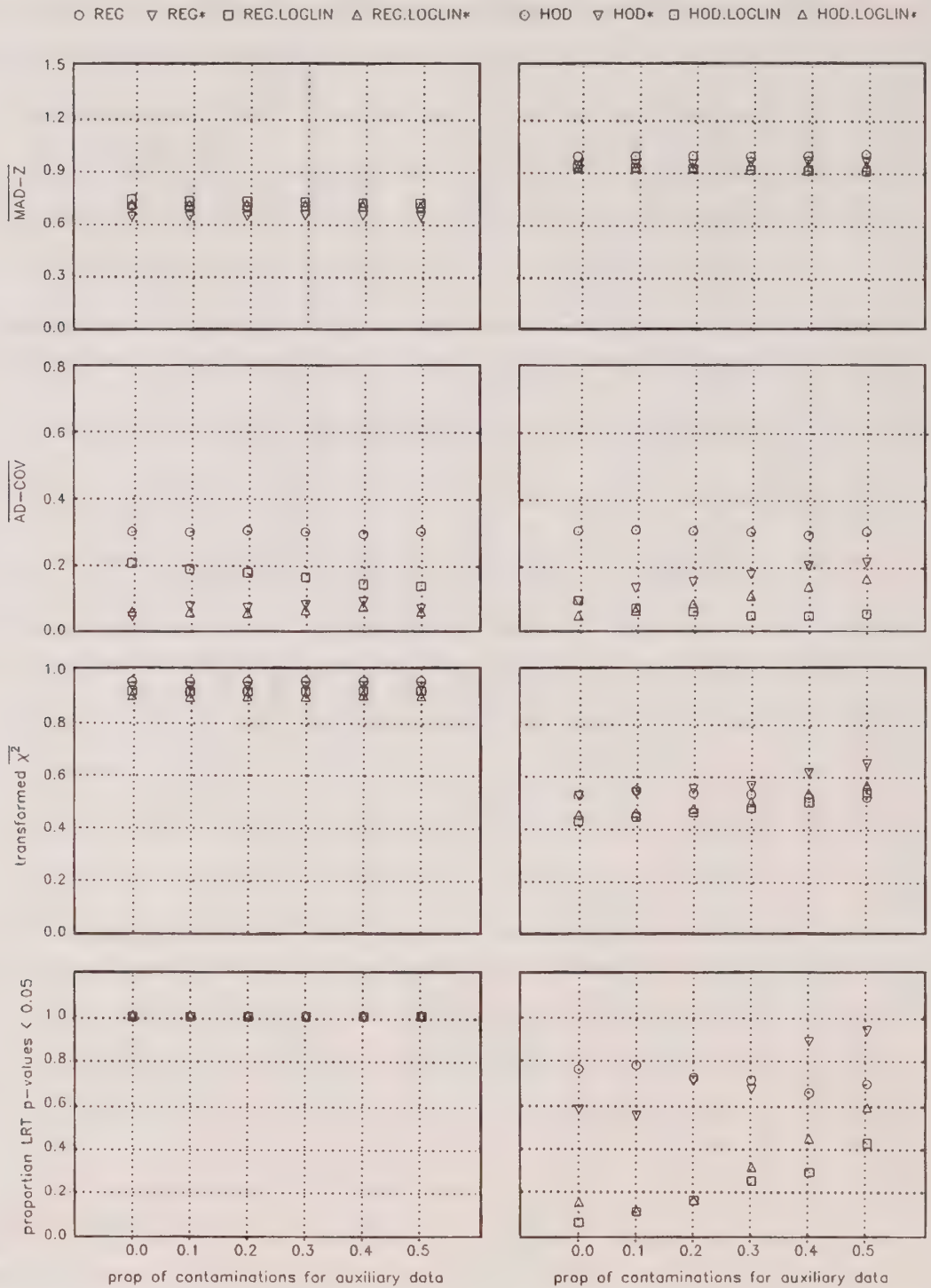


Figure 5. Comparison of statistical matching methods as the proportion of log normal contaminations varies for the auxiliary data file C ($\rho_{Y,Z|X} = .4$ before contamination and files A and B have no log-normal contamination)

In the hot deck family of methods with auxiliary information, the two methods with categorical constraints, namely, HOD.LOGLIN and HOD.LOGLIN* show vary favourable performance at the aggregate level (*i.e.* with respect to transformed χ^2 and LRT) for all types of underlying populations; see Figures 2 to 5. In this case, the method HOD.LOGLIN generally outperforms HOD.LOGLIN*. Next we consider unit level measures. For symmetric and skewed populations (Figures 2 and 3), generally speaking, HOD* and HOD.LOGLIN* perform very similarly to each other and somewhat better than HOD.LOGLIN but slightly worse than REG*. However, for proxy auxiliary data (Figures 4 and 5) with respect to the AD-Cov measure, the method HOD.LOGLIN could be better or worse than the HOD.LOGLIN* and the REG* methods, and is more often better in the case of proxy data with log normal contaminations. Also HOD.LOGLIN in the case of proxy data tends to have fairly robust behaviour with respect to all four evaluation measures. Thus, in the hot deck family, based on overall performance, HOD.LOGLIN* can be recommended. However, in practice, HOD.LOGLIN may be preferable as a compromise because it performs moderately well at the unit level, extremely well at the aggregate level, is computationally much less demanding and shows robustness with respect to the proxy auxiliary data. Furthermore, HOD.LOGLIN does not require micro-level auxiliary information.

7.3 Miscellaneous Observations

In this subsection we describe separately some interesting findings, the corresponding empirical results for some of which are not included here, but are presented in Singh *et al.* (1990).

(i) Shrinkage to the Mean

An important and consistent finding was that matching methods in the regression family do not perform well with respect to the categorical measures. This can be explained by shrinkage towards the mean; that is, the matched Z -values are more tightly distributed about their mean than are the suppressed true Z -values. This is displayed in Figure 1 which shows the difference between the marginal histograms of matched and suppressed Z -values for various matching methods.

The positive differences for REG and REG* near the centre indicate that there are more Z -values in that region in the matched file than in the suppressed file. The very large negative observations at the extreme points of this plot are associated with open ended intervals, and it seems quite likely that had these intervals been broken down into several smaller intervals the plot would have shown several smaller negative numbers in the extreme tails, so that the interpretation of the plot should be that these methods are

putting too many Z -values at the centre of the distribution at the expense of the extreme tails.

Figure 1 also shows shrinkage towards the mean for the REG.LOGLIN and REG.LOGLIN* methods. However, in this case the shrinkage is limited by the categorical constraints so that, while we still see that the tails of the Z -distribution of the matched file are too short, the displaced values are now not going to the centre of the distribution, but only to the partition boundary points which act like walls. The large positive values to either side of the central boundary point can be explained similarly if one bears in mind that what this plot is showing is actually an average of differences of histograms over 100 independent simulations. It seems reasonable that if we were to examine each of the 100 differences of histograms individually we would sometimes see a large positive value just to the left of the central boundary point, and sometimes just to the right, but never both at the same time.

Figure 1 also shows that shrinkage to the mean and boundary effects are not serious for methods in the hot deck family.

(ii) (Y,Z) vs (X,Y,Z) Auxiliary Information

Although only results based on (Y,Z) auxiliary information were presented in this paper, (X,Y,Z) auxiliary information was also considered as part of the simulation study as mentioned in Section 6. An interesting finding was that for the HOD.LOGLIN and HOD.LOGLIN* methods, the use of (Y,Z) auxiliary information leads, in general, to somewhat better performance at the aggregate level than the use of (X,Y,Z) information. This does not seem to be the case with the HOD* method. This phenomenon is probably due to instability in the estimation of (X^*, Y^*, Z^*) factor effects used in the categorical constraints on account of insufficient size of auxiliary data. An implication is that the true values were probably close to zero and so taking them as zero leads to better results. This suggests that the impact of different sample sizes on performance of matching methods should be considered, if possible, in future investigations. The above consideration also suggests an interesting new class of methods which would combine (X,Y,Z) micro-level auxiliary information for finding Z_{int} values along with the derived (Y^*, Z^*) categorical distribution only from file C for imposing constraints. These methods were, however, not included in the present study.

(iii) Comparison of Different Versions of Hot Deck Methods

In all the matching methods considered, except HOD and HOD.LOGLIN, the second step for finding Z_{match} consists of using hot deck imputation in which (X,Z) -distance is employed. For the remaining two, X -distance was considered. Some other options (for methods other than

HOD and HOD.LOGLIN), consist of using Z -distance or (X, Y, Z) -distance. For the latter, Y_{int} would have to be added first to file B. This was included in the original simulation study, although empirical results are not reported here. It was found that there is generally no difference though, for REG and REG* methods, (X, Z) -distance sometimes showed superior performance with respect to the AD-Cov measure. This is the reason for our choice of (X, Z) -distance in the methods considered here. However, in practice, it may be preferable to use Z -distance with hot deck matching methods because of computational convenience.

Further, it should be noted that for HOD and HOD.LOGLIN methods, there is the option of using random or rank in Step II instead of X -distance. In hot deck rank, records from files A and B are ranked separately according to the value of X , and then are matched based on ranks. This was proposed by G. Rowe for the SPSS application mentioned in the introduction. Clearly, this method is suitable for univariate X only. An advantage of ranking is that there will not be one record from file B acting as donor for many records from file A. The above three versions of hot deck were included in the Monte Carlo study although results for X -distance only are reported here. It was found that it generally does not make much difference which version is used. The choice of X -distance was made for HOD and HOD.LOGLIN because it was consistent with the hot deck distance version used for other methods. In practice the hot deck random version would be least demanding computationally; however, in a real application we would not know how much might be lost by using random matching instead of ranking or distance, and we would probably want to use as much information as would be feasible.

8. CONCLUDING REMARKS

In this paper, the problem of using auxiliary information in statistical matching was considered. The two main methods previously proposed are due to Rubin (1986) and Paass (1986), versions of which were denoted by REG* and HOD*. Some modifications of these methods, denoted by REG.LOGLIN* and HOD.LOGLIN*, were proposed by imposing categorical constraints derived from auxiliary information. These would reduce to REG.LOGLIN and HOD.LOGLIN if only categorical auxiliary information is available or useable. In the absence of auxiliary information, the usual methods of imputation, REG and HOD would be used. An empirical study was conducted to evaluate performance of the above eight methods with respect to four evaluation measures (two at the unit level, and two at the aggregate level). It was found that for the case of no auxiliary information, the HOD method is preferable. The case of auxiliary information is, however, more complex. If only unit level evaluation measures are deemed important, then the REG*

method is recommended. If aggregate level measures are also considered important then if there is nonproxy auxiliary data HOD.LOGLIN* is recommended. As an alternative, a good compromise would be HOD.LOGLIN if computational burden is an important consideration or if proxy auxiliary data is believed to be present. If unit level measures are less important or are not of interest (this may often be the case because the matched data would generally be presented in tabular forms in practice), then HOD.LOGLIN would be recommended. With both HOD and HOD.LOGLIN methods, the similar performances of distance, random and rank versions might suggest the use of random versions in practice in view of its computational simplicity.

It may be remarked that we did not consider the fully iterative version of Paass's method. It would be interesting to find out in future investigations how this might perform. Another point that requires investigation is the implementation of categorical constraints with many variables. The application of the raking algorithm may be computationally prohibitive. In this connection, the results of Paass (1989) are expected to be useful.

In the present study we did not, due to limitations of computing, systematically vary the accuracy of the auxiliary data source; that is, we did not vary the size of the file C. We also did not vary the size of the files A or B. An interesting question that might have been addressed is how the performance of various methods might be affected by the size of these files.

Finally, it should be pointed out that although the results of this study are based on synthetic data (which was necessary to produce various scenarios mimicking real data), it is believed that the results would be relevant for real applications. Clearly, it would be interesting and useful to carry out a simulation study with real data to check whether the findings continue to hold and to see what sorts of substantive impact the biases in the joint distribution of the matched file have. A related question is how to account for such biases in inferences based on the matched file; that is, how to produce measures of uncertainty for parameter estimates from the matched file that reflect not only the variability within the matched file, but also the uncertainty inherent in the matching procedure itself. Although we are unable to answer this question, it is clear that matching procedures using auxiliary information would enhance the overall utility of the matched file. These and some other related questions will be investigated in the future.

ACKNOWLEDGEMENT

The authors would like to thank J. Armstrong, G. Gray, G. Hole, D. Royce and M. Wolfson for helpful comments. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

REFERENCES

- ARMSTRONG, J. (1989). An evaluation of statistical matching methods. Methodology Branch Working Paper, BSMD, 90-003E. Statistics Canada.
- BARR, R.S., and TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Surveys. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.
- BARR, R.S., and TURNER, J.S. (1990). Quality issues and evidence in statistical file merging. In *Data Quality Control: Theory and Pragmatics* (Eds. G.E. Liepins and V.R.R. Uppuluri). New York: Marcel Dekker, 245-313.
- BARR, R.S., STEWART, W.H., and TURNER, J.S. (1981). An empirical evaluation of statistical matching methodologies. Technical report, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.
- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.
- BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.
- BUDD, E.C., and RADNER, D.B. (1969). The OBE size distribution series: methods and tentative results for 1964. *American Economic Review, Papers and Proceedings*, LIX, 435-449.
- COHEN, M.L. (1991). Statistical matching and microsimulation models. In *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Volume II, Technical Papers, (Eds. C.F. Citro and E.A. Hanushek). Washington, D.C.: National Academy Press, 62-85.
- FELLEGI, I.P. (1977). Discussion paper. *Proceedings of the Section on Social Statistics, American Statistical Association*, 762-764.
- FORD, B.L. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, (Vol. 2), (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 185-207.
- KADANE, J.B. (1978). Some statistical problems in merging data files. In *1978 Compendium of Tax Research*. Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.
- PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy* (Eds. G.H. Orcutt, J. Merz and H. Quinke). Amsterdam: Elsevier Science.
- PAASS, G. (1989). Stochastic generation of a synthetic sample from marginal information. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 431-445.
- PAASS, G., and WAUSCHKUHN, U. (1980). Experimentelle erprobung und vergleichende Bewertung statistischer Matchverfahren. Internal report, IPES.80.201, St. Augustin, *Gesellschaft für Mathematik und Datenverarbeitung*.
- PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or Domains). *International Statistical Review*, 48, 3-18.
- RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91-102.
- RODGERS, W.L., and DeVOL, E. (1982). An evaluation of statistical matching. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-132.
- RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87-94.
- RUGGLES, N., RUGGLES, R., and WOLFF, E. (1977). Merging microdata: Rationale, practice and testing. *Annals of Economic and Social Measurement*, 6, 407-428.
- SCHEUREN, F.J. (1989). Comment on Wolfson *et al.* (1989). *Survey of Current Business*, 69, 40-41.
- SIMS, C.A. (1972). Comment on Okner (1972). *Annals of Economic and Social Measurement*, 1, 343-345.
- SIMS, C.A. (1978). Comment on Kadane (1978). In *1978 Compendium of Tax Research*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.
- SINGH, A.C. (1988). Log-linear imputation. Methodology Branch Working Paper, SSMD, 88-029E, Statistics Canada; also published in *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 118-132.
- SINGH, A.C., ARMSTRONG, J.B., and LEMAÎTRE, G.E. (1988). Statistical matching using log linear imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 672-677.
- SINGH, A.C., MANTEL, H., KINACK, M., and ROWE, G. (1990). On methods of statistical matching with and without auxiliary information: Some modifications and an empirical evaluation. Methodology Branch Working Paper, SSMD, 90-016E. Statistics Canada.
- U.S. DEPARTMENT OF COMMERCE (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, Washington, D.C.: Federal Committee on Statistical Methodology.
- WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B., and ROWE, G. (1987). The social policy simulation database: an example of survey and administrative data integration. *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada, Ottawa, (Eds. J.W. Coombs and M.P. Singh), 201-229; another version published in *Survey of Current Business* (1989), 69, 36-40.

A Framework for Measuring and Reducing Nonresponse in Surveys

MICHAEL A. HIDIROGLOU, J. DOUGLAS DREW,
and GERALD B. GRAY¹

ABSTRACT

The need for standards introduced for the gathering and reporting of information on nonresponse across surveys within a statistical agency is discussed. Standards being adopted at Statistics Canada are then described. Measures to reduce nonresponse undertaken at different stages in the design of surveys at Statistics Canada that have a bearing on nonresponse are described. These points are illustrated by examining nonresponse experiences for two major surveys at Statistics Canada.

KEY WORDS: Nonresponse rates; Incentives; Follow-ups; Data collection.

1. INTRODUCTION

National agencies such as Statistics Canada conduct a large number of different surveys every year. These vary in their subject matter, units of response, periodicity, sample design and collection methodologies. They also have varied experiences with respect to the nonresponse incurred. There is a need for agency-wide standards for the gathering and reporting of information on response and nonresponse. If they are sufficiently flexible to accommodate the requirements of the variety of surveys that are conducted, it is logical to have standard definitions. A distinction needs to be made though between standard definitions and standards of acceptable levels of different components of nonresponse to surveys. It is the former and not the latter that is under discussion.

There are major differences between surveys that result in different levels of nonresponse achieved; for example, longitudinal and cross-sectional surveys face somewhat different missing data problems. Standard definitions can provide a common lexicon that will help in isolating and understanding better the differences. A common lexicon helps in the ongoing analysis of trends in nonresponse. Information on survey response and nonresponse can serve multiple purposes, such as the potential for nonresponse biases, pointing to weak areas that need to be strengthened in future rounds of the survey. They provide measures of frame coverage, for developing methods to compensate for and to reduce nonresponse. They also give an important input to survey design, collection methodologies, evaluation of data quality and operations for different surveys.

Nonresponse rates can be defined differently, depending on whether they are used to diagnose sampling activities,

data collection activities or to analyze published data. For example, in the case of sampling requirements, the unit for which nonresponse is measured ought to be the sampled unit. Correspondingly, for data collection activities, the unit of measure for computing nonresponse would be based on the unit of response. It should be noted that for business surveys there is often not a one-to-one correspondence between sampled units and units of response (*e.g.*, the sampled unit may be the head office and the unit of response is its branches). For published data, the measure of nonresponse could be weighted size measures or weighted key variables to estimate the contribution of nonrespondents to the key aggregates. In business surveys, such measures can be important because of the skewed populations where a few units contribute to a disproportionately large share of the estimate.

Breakdowns of the nonresponse rates should be available at pre-determined geographical levels, industrial and size levels and combinations of it. If possible, the reasons for nonresponse also should be available *e.g.*, unable to contact, refusal *etc.* These can be used to produce diagnostics to establish causes of nonresponse. If the data are collected by using interviewers located throughout nationwide regional offices, then nonresponse rates by interviewers within each regional office and nonresponse rates aggregated by regional office can be used as measures of operational performance. Questionnaire item nonresponse rates can be used to point to questions that need to be rethought in terms of wording or data availability.

This paper deals with total nonresponse, where nonresponse occurs at the level of the unit for which data are being collected. It does not deal with partial nonresponse, where the respondent provides usable information for some items but not for others. We start with a conceptual

¹ Michael A. Hidiroglou, Business Survey Methods Division; J. Douglas Drew, Household Surveys Division; Gerald B. Gray, Social Survey Methods Division, Statistics Canada.

framework for the definition of response and nonresponse that is suitable for both business and social surveys. The next section is devoted to general causes of nonresponse and to means for reducing nonresponse. Finally, we look at the experiences with nonresponse for two major surveys conducted at Statistics Canada.

2. DEFINITIONS OF NONRESPONSE RATES

Nonresponse rates and their complements, response rates, are defined as ratios of variables that represent a given category of response/nonresponse in some domain of interest. The important variable may be a simple count or it may be weighted by some factor. It may be the sample weights of the unit or the unit's expected contribution to the estimate of some major statistic of the survey. Figure 1 represents a conceptual framework developed by Drew and Gray (1991) for classifying sampled units in a survey into responding, nonresponding and out-of-scope units. The hierarchical representation is similar to one initially proposed by Platek and Gray (1986). The framework has been evaluated for several business and social surveys at Statistics Canada. The agency has adopted these standards

for the gathering and reporting of information on nonresponse. Starting with the 1993 reference year, several major surveys will be required to report detailed nonresponse using the standard definitions. A data base of nonresponse rates will be maintained for use in agency-wide monitoring and analysis of trends in nonresponse.

We begin with the **Total number of Units** (weighted or unweighted). The total number of units consists of those that are thought to belong to the survey of interest before the survey process begins. The total (Box 1 in Figure 1) is broken down into two main categories: resolved (Box 2) and unresolved (Box 3) units. **Resolved Units** are those whose status as belonging or not belonging to the target universe is known by the cutoff date of the survey data collection. For some surveys all units can be resolved. For other surveys, it is either impossible or impractical to resolve all units. For example, in a telephone survey there are telephone numbers that ring but do not correspond to working numbers. Without checking the status of each so-called ring-no-answer case with the telephone company, there is no way to determine whether such a number represents a working number. Similarly for a survey with mail collection, without a follow-up of units not returning a questionnaire, it may not be known which units are

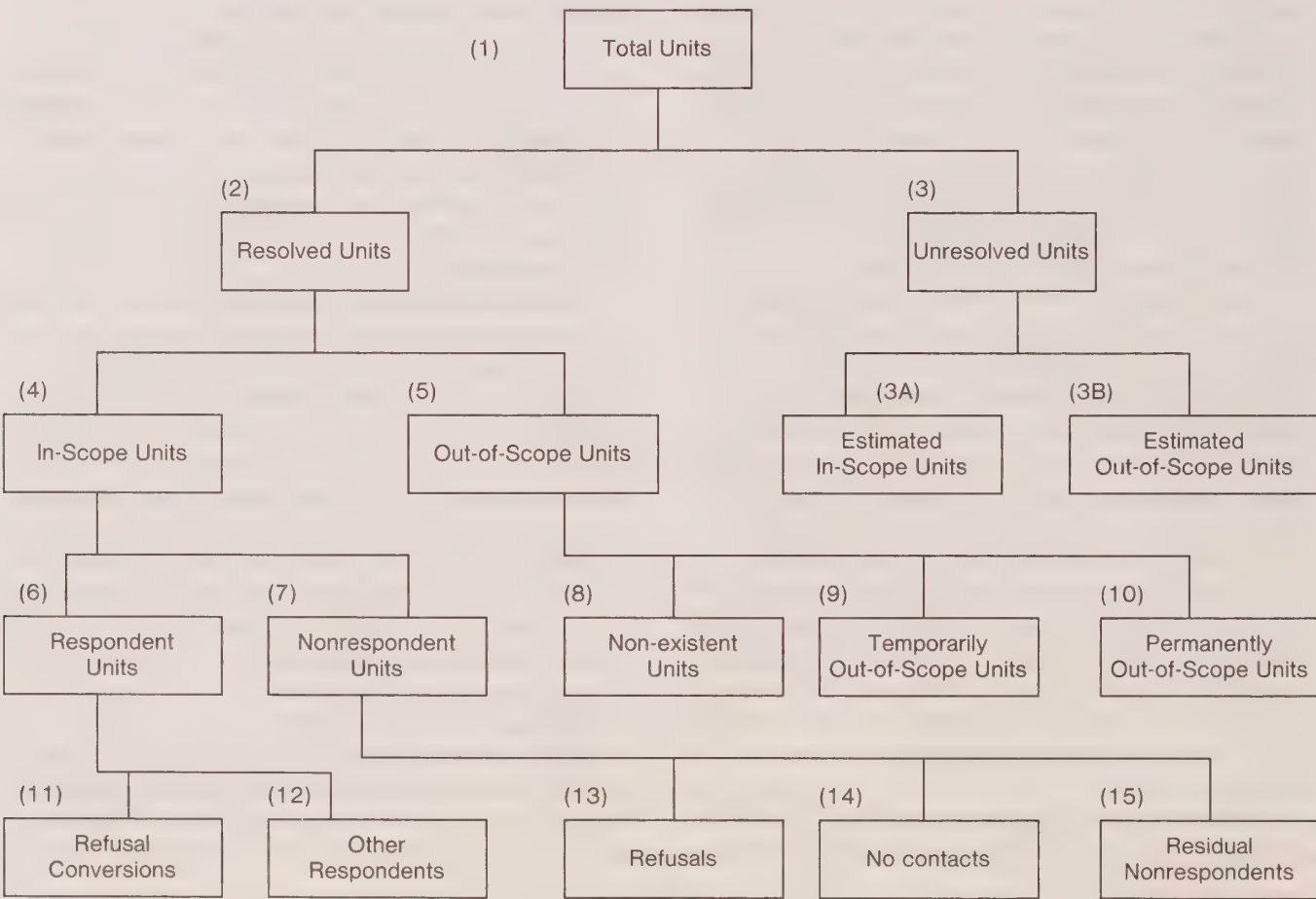


Figure 1. Respondent/Nonrespondent Components at the Data Collection Phase

out-of-scope (e.g., the unit no longer exists, or it exists but it is out-of-scope), versus those that are in-scope and should have responded. **Unresolved Units** are units whose status cannot be determined by the end of data collection for the survey. The number of Unresolved Units may be broken down into **Estimated In-Scope Units** and **Estimated Out-of-Scope Units** by apportioning the number in the same ratio as for the Resolved Units for example. A **Resolved Rate** may then be defined as the *ratio of the number of Resolved Units to the Total number of Units*. The two components of the Resolved Units, i.e. In-Scope (Box 4) and Out-of-Scope (Box 5), lead to two complementary rates: the **In-Scope Rate**, defined as the *ratio of the number of In-Scope Units to the number of Resolved Units* and its complement, the **Out-of-Scope Rate**.

The Out-of-Scope Units (Box 5) may be split up into as many as three categories, some of which may not be applicable to a particular survey. These include Non-existent (Box 8), Temporarily Out-of-Scope (Box 9) and Permanently Out-of-Scope (Box 10) Units. The **Non-existent Units** include business deaths, that is, companies that have gone out of business, and dwellings that have been demolished. For recurring surveys, once it is determined that a unit is non-existent, it is excluded from data collection on future survey occasions. The **Temporarily Out-of-Scope Units** are units that were Out-of-Scope at the time of the survey, but which might be in-scope later. Hence, units can be temporarily Out-of-Scope even for single occasion surveys. For recurring surveys, it is necessary to recontact temporarily out of scope cases periodically in case their status has changed. Examples include businesses that are inactive due to seasonal factors, seasonal dwellings whose occupants have a usual place of residence elsewhere, and vacant dwellings. The **Permanently Out-of-Scope Units** result from improper classification on the frame because of changes in the classification since the frame was last updated. These cases may be screened out during the first stage of response. The Out-of-Scope Rate may be split up into three component rates: the **Non-existent Rate**, defined as the ratio of the number of Non-existent Units to the number of Resolved Units. The **Temporarily Out-of-Scope Rate** and **Permanently Out-of-Scope Rates** rates are similarly defined.

The In-scope Units (Box 4) may be broken down into Respondent (Box 6) and Nonrespondent Units (Box 7). The **Respondent Units** include in-scope units that have responded by the cutoff date for the data collection and have provided "usable information". The notion of "usable information" applies to respondents who provide only partial information. A threshold is needed in terms of level of completion of the questionnaire, below which units are considered nonrespondents. The **Response Rate** may be defined in different ways, depending upon the intended analysis. We prefer to define it as ratio of the

number of Respondent Units to the number of In-Scope and Unresolved Units. This ratio is a conservative measure of the quality of the frame and the data collection procedures, since some Unresolved Units may be Out-of-Scope. An alternative definition would include only the number of In-Scope Units in the denominator. That rate, a conditional response rate given the known status of the units in the sample, measures the efficiency of the data collection procedure alone. The **Nonrespondent Units** (Box 7) are the remainder of the In-Scope Units. The **Nonresponse Rate** is defined as the complement of the Response Rate. It is the *ratio of the number of Nonrespondent and Unresolved Units to the number of In-Scope and Unresolved Units*. Alternative definitions omit the Unresolved Units in the numerator and denominator or apportion the Unresolved Units between estimated numbers of In-scope and out-of-scope units.

To determine the effort needed to convert Refusals to Respondents at the data collection stage, the Respondent Units are divided between Refusal Conversions (Box 11) and Other Respondents (Box 12). The **Refusal Conversions** are those who refuse initially in the current or previous collection period, and are successfully converted to be respondents because of follow-up interviews. The **Refusal Conversion Rate** is a measure of the success in converting refusals to respondents. Instead of being merely a component of the Response Rate with the same denominator, the **Refusal Conversion Rate** is defined as the *ratio of the number of Refusal Conversions to the number of Refusals and Refusal Conversions*. For completeness, we label respondents who were not refusal conversions as **Other Respondents**.

Finally, the Nonrespondent Units (Box 7) may be broken down into three components; viz. Refusals (Box 13), No Contacts (Box 14) and all remaining categories, that is, the Residual Nonrespondents (Box 15). The **Refusals** are nonresponding units that have been contacted but refuse to participate in the survey. The **Refusal Rate** is defined as the *ratio of the number of Refusals to the number of In-Scope Units*. The **No Contacts** are in-scope units that cannot be contacted. For social surveys, these include dwellings whose occupants were temporarily absent and households where no one was at home when interviews were attempted. The occupancy status of such dwellings is determined through observation, or where applicable by speaking to building superintendents. For business surveys, these include telephone respondents who cannot be reached, and mail nonrespondents known to be in-scope, but who were not contacted as part of any nonresponse follow-up. The **No-Contact Rate** is defined as the *ratio of number of No-Contacts and Unresolved Units to the number of In-Scope and Unresolved Units*. The **Residual Nonrespondents** include units that did not respond due to special conditions (for example, language problems, or inaccessibility) as well as respondents who provided no usable information. Special conditions also include in-scope units for which interviews

were not attempted. This is to avoid unwanted overlap between samples for different surveys, as a measure to prevent undue respondent burden. While these latter units differ from other nonresponse in that interviews are not attempted, it is important that they be considered as non-respondents in deriving nonresponse adjustment factors at the estimation stage. The **Residual Nonrespondent Rate** is the *ratio of the number of Residual Nonrespondents to the number of In-Scope units*.

The rates defined above are at the unit level. Clearly, rates can also be defined at the item level, that is, for individual items on the questionnaire. Typically an item tends to be completed, missing, or in error as detected during editing. Hence, at an item level one can define a response rate, a missing rate, and an edit failure rate. If the missing or edit failures are imputed, one can define an imputation rate. These rates can be defined for unit respondents only, which would generally be preferable if unit nonresponse is treated by reweighting. Alternatively, unit nonrespondents can be included in both the numerators and denominators for the rates, which would be preferable in cases where unit nonresponse is treated by imputation.

We apply the definitions of nonresponse as provided above to several business and social surveys. Table 1 presents annual average nonresponse rates for several Statistics Canada surveys.

Nonresponse rates are highest for three of the social surveys and stem from: (i) the sensitivity of income as a subject matter in the case of the Survey of Consumer Finances, (ii) the respondent burden due to the length of the interview in the case of the Family Expenditure Survey, and (iii) the combination of inexperienced interviewers, telephone survey methodology and nonproxy reporting for the General Social Survey. Nonresponse is very low for the LFS because this is a long-standing flagship survey where many steps are taken to keep nonresponse low. Nonresponse rates are low for the business surveys in Table 1. Some initiatives were undertaken during the recent business survey redesign program, at reducing nonresponse.

3. FACTORS AFFECTING NONRESPONSE

Several survey design factors impact on response and nonresponse. In this section, we begin by briefly examining the influence of the frame and sample design. We follow with a more in depth examination of data collection, in terms of its organization, interviewer training, technology, mode of primary collection, and questionnaire design. Also, methods used for follow-up of nonresponse and edit failures, and use of administrative data to replace direct collection are considered.

Table 1
Response Rate Components for Selected Surveys Data Collection Stage (Rates in %)

	COMPUTATION	FAMEX	ASM	GSS	SCF	LFS	RTS
Resolved rate	(2)/(1)	100.0	100.0	98.1	100.0	100.0	95.8
In-Scope rate	(4)/(2)	92.0	95.3	51.2	86.3	85.1	97.0
Response rate	(6)/[(3)+(4)]	72.9	92.8	75.9	73.9	94.4	94.0
Refusal Conversion rate	(11)/[(11)+(13)]	N.A.	N.A.	26.7	N.A.	N.A.	N.A.
Nonresponse rate	[(7)+(3)]/[(3)+(4)]	27.1	7.2	24.1	26.1	5.6	6.0
Refusal rate	(13)/(4)	16.2	7.2	13.2	23.7	1.5	1.7
No-Contact rate	[(14)+(3)]/[(3)+(4)]	5.1	0.0	5.9	2.3	3.6	4.3
Residual nonresponse rate	(15)/(4)	5.8	0.0	5.8	0.0	0.4	0.0
Out-of-Scope rate	(5)/(2)	8.0	4.7	48.8	13.7	14.9	3.0
Non-Existent rate	(8)/(2)	0.8	2.5	0.0	0.3	0.3	2.3
Temporarily Out-of-Scope rate	(9)/(2)	7.1	1.2	0.0	13.4	14.6	0.5
Permanently Out-of-Scope rate	(10)/(2)	0.0	1.0	48.8	0.0	0.0	0.3

FAMEX: Family Expenditure Survey (1990).

ASM: Annual Survey of Manufactures (1989).

GSS: General Social Survey Cycle 5 (January-March 1990).

SCF: Survey of Consumer Finances (1991).

LFS: Labour Force Survey (1990).

RTS: Retail Trade Survey (December 1990).

3.1 Frame

Duplication or overcoverage on a frame can be irritating and lead to nonresponse if there are no procedures for unduplication, or if the procedures are not always successful. For business surveys, accurate classification information is essential if the survey is industry specific or uses industry specific questionnaires. For example, if a sampled business receives a questionnaire not pertaining to its industrial activity, it is unlikely to respond. Accurate information on the coverage of complexly structured businesses is necessary to provide respondents with a good description of the required geographical and/or industrial information. Similarly, information on contact persons within the business is needed to establish good reporting arrangements with the respondent. Inaccuracies in the contact information will cause delays in getting the required data. Inaccurate coverage description will result in improper or incomplete data being provided by the respondent.

The samples for Business surveys at Statistics Canada are drawn from a file known as the Business Register. It is a list frame that contains relevant information for selecting and contacting samples of business respondents. It has recently been redesigned using a comprehensive model reflecting the real-world complexity of business respondents. The processes incorporated in the Business Register minimize the impact of the above causes for nonresponse. Duplication is kept to a minimum by continually linking the changes that are occurring to existing units on the Business Register. These changes include births, amalgamations, splits and mergers of business respondents. Several events can signal changes to the structure of large businesses, including different administrative sources and direct survey feedback. These signals trigger a "profiling" action, *i.e.* contact with the business to redefine its structure. In the absence of signals, structures are profiled on a periodic basis, at a frequency depending upon their significance and their propensity to change. The profiling exercise gathers the necessary information to update the model. More details of the required actions are provided by Colledge (1989). The source of updates is a combination of administrative updates, profile updates and direct survey feedback. Contact, coverage and questionnaire type is kept up-to-date for each sampled unit by setting up and maintaining a computerized collection system for sampled businesses for each survey of interest. The resulting collection units are automatically built and kept up-to-date using well defined rules that vary from survey to survey. The questionnaire type takes account of factors such as: the periodicity of data collection, industrial classification, any seasonal considerations for sub-annual surveys, and fiscal year ends for annual surveys. Automatic maintenance of these collection units is carried out using a wide range of updates to the Business Register. These updates encompass activity status (live, dead, seasonal), name, address and telephone changes as well as structural changes to the surveyed unit.

The adequacy of the frame plays a similar role for social surveys in reducing nonresponse. The frame in combination with the sample design and collection procedures is important: in ensuring manageable interviewer workloads, in providing information to facilitate contact of respondents by interviewers, and in preventing unwanted overlaps in the sample across surveys. The Labour Force Survey (LFS) serves as the main vehicle for the conduct of social surveys based on area sampling. Presently, most other social surveys are supplements to the LFS, that are administered through add on questions to LFS respondents. Some surveys, due to the length of the interview or sensitivity of the subject matter are not suitable as supplements. Instead, they are based on separate samples of households drawn from the LFS frame and design.

The LFS is based for the most part on an area frame, and initial contact with sampled households is generally by face-to-face interview. The efficiency of the area frame deteriorates over time; dwelling counts for the sampling units used to determine the selection probabilities of the sampling units and interval of sampling become out-of-date. This makes it harder to plan and maintain manageable interviewer workloads. The principal mechanism for keeping the area frame up-to-date is a sample redesign following each decennial census of population. Other measures have included *ad hoc* frame updating restricted to high growth areas identified by the mid-decade census. Another measure taken in the 1981 redesign was the creation of so-called buffer strata on the outskirts of large urban centres. This involved a simple design that could be readily updated without affecting the remainder of the frame in the event that growth of the urban centre reached out into the buffer zone. To prevent interviewer workloads from becoming unwieldy when units experiencing large growth enter the sample, sub-sampling is done. For cases of extreme growth, area sub-sampling is resorted to, in which the areal unit is sub-divided into new units, a sub-sample of which is selected. If the growth is not too high, the original sample unit is retained. The rate of sampling is modified to reduce the number of dwellings selected to the point where it no longer poses a problem in terms of the interviewer's workload.

Besides the area frame, a list frame of apartment buildings is used by the LFS in larger cities. This list is kept current using information on building permits. To facilitate contact with sampled dwellings in the apartment sample, telephone numbers are obtained, where possible, by matching address information to telephone company files. Supplying interviewers with telephone numbers in this fashion has proven useful since it gives them an additional means of contacting selected dwellings that are difficult to access due to security systems, or where it is difficult to find people at home. Since the introduction of this procedure, while the nonresponse rate for the apartment frame remains higher than that for the area frame, the gap has narrowed from 8.6% to 6.2%. An alternative

to the area frame used by a few social surveys is a telephone frame. Sampling is based on Random Digit Dialing of numbers within "banks" of numbers containing working residential numbers. The banks are updated using files purchased from telephone companies. To prevent undue respondent burden, telephone numbers of households currently or recently in the LFS or other surveys using the area frame are excluded from the telephone surveys.

The LFS is currently being redesigned. Consideration is being given to adopting an address register as a list frame in urban areas. An address register of residential dwellings was created as a coverage improvement tool in the 1991 Census, and is being updated to reflect the Census enumeration of dwellings (Swain *et al.* 1992). Ways of updating the address register on an ongoing basis using administrative records or information from the postal service, and using it as a frame for social surveys are currently under study. An address register based frame should impact positively on field operations and nonresponse. Telephone numbers will be available for up to 70% of dwellings as a tool for interviewers to facilitate contacting households. Due to its regular updating, the sample can be designed to have good control on interviewer workloads, without having to resort to measures such as sub-sampling as are required under the area frame. Additionally, for the redesign, it is planned to build in mechanisms for both area and list frames to track all dwellings that are selected for Statistics Canada surveys.

3.2 Sample Design

The sample size for a survey is arrived at by taking into account budgets, survey objectives and desired level of reliability for key variables for the primary domains of interest. The overall sample size and survey design strategy should also allow for follow-up of non-responding units. In Section 4, we illustrate this point for the recently redesigned Monthly Wholesale and Retail Trade Surveys at Statistics Canada.

Business and Agricultural Surveys are stratified by a number of key variables including the size of the units. Because of the highly skewed nature of the distribution of key variables in the population, the size stratification results in a take-all and a number of take-some strata. Units in the take-all stratum cannot be rotated out of the sample, unless they become smaller in size over time. Optimum sampling plans that minimize the overall sample size for given levels of reliability may require too many units in the take-all stratum. To minimize response burden, some surveys restrict the number of take-all units; for example, the National Farm Survey (Julien and Maranda, 1990). Another means under consideration to reduce the response burden among the large units is the integration of questionnaires and/or data collection for several surveys. This implies that only distinct statistical data need to be collected for the different surveys.

Response burden among the smaller units can be reduced by periodic rotation of sampled units. However, rotation of units increases the cost of the survey because of additional sample maintenance, additional training of interviewers and difficulties in grooming new units to provide data. Partial rotation of sampled units at some fixed rate is undertaken as a compromise between 100% rotation which is very expensive and gives poor estimates of change, versus no rotation at all which would result in an unacceptable distribution of response burden. The rotation schemes keep a unit in the sample for a given period of time, after which the unit would be ineligible for reselection by the same survey for a minimum period. Surveys using such a scheme include: the Survey of Employment, Payrolls and Hours (with rotation of approximately 1/12th of the take-some units of the sample every month), the Monthly Wholesale and Retail Trade Survey (with rotation of approximately 1/24th of the smaller sized units every month), and the Labour Force Survey (with rotation of 1/6th of the sample every month). Another way to reduce response burden for individual units of Business and Agricultural Surveys is to minimize the overlap between surveys. This can be accomplished using a technique known as synchronized sampling. This technique attaches a permanent random number between 0 and 1 to each unit in the population. Different surveys are then allotted subsets of the interval (0,1) and all units whose random number falls within a survey's allotted subset are selected for that survey.

One of the objectives in the redesign of the Labour Force Survey to be introduced in 1995-1996 is to achieve a general household survey vehicle. Several new recurring social surveys are scheduled to start up in the mid-1990's, including a longitudinal survey of labour and income dynamics, and a health survey. The LFS redesign will consider not only LFS requirements, but requirements of these other surveys. Elements of the general survey orientation will include a common frame and similar sample designs with general purpose stratification. It will also feature co-ordinated sampling with overlap of selected primary sampling units (PSU's) to permit common interviewers across surveys. Unduplication of samples of dwellings between surveys to avoid respondent burden will also be carried out.

3.3 Data Collection Procedures

While all facets of the survey design can influence the survey response rates, data collection procedures and operations have the most direct and important bearing. In this section we examine the data collection procedures for business and social surveys, and the impact that factors such as the organization, the interviewer, mode of collection, technology, follow-up strategies, and response incentives have on nonresponse.

3.3.1 Organization of Data Collection

Data for business surveys are collected primarily through mail surveys with telephone follow-up. Before the mid-1980's, the collection and editing of business survey data was carried out principally in the subject matter divisions of Statistics Canada at its Head Office. This resulted in over seventy percent of the staff in these divisions being assigned to the processing of survey data. For many business surveys, regional offices had the responsibility of collecting data for nonrespondents to the surveys. During the mid-1980's, it was recognized that better use of Head Office and regional office resources could be made by a shift in the organization of data collection. The shift resulted in the concentration of collection and data capture activities within one division at Head Office specializing in the collection of annual data, and the regionalization of data collection for sub-annual surveys to the regional offices. The benefits of this reorganization were as follows: (i) operational resources could be used more effectively, (ii) the division of resources between the Head Office and regional offices could be better allocated, (iii) the increasing complexity of data collection could be handled by groups specialized in this activity, and could more readily exploit technical innovations and movement towards more integrated collection procedures, (iv) regional offices could establish "warm" contacts with the potential respondents on account of their geographical proximity to them, and (v) regional offices could offer services to users that would enhance Statistics Canada's presence among the potential responding units. All this helped in reducing the nonresponse rates.

Data for the social surveys are collected through a combination of face-to-face and telephone interviews. The monthly Labour Force Survey and most other social surveys conducted by Statistics Canada use a dispersed field force of approximately 1,000 interviewers across the country. The interviewers do a mixture of telephone interviewing from their homes and face-to-face interviewing. They are supervised by 100 senior interviewers. Project managers located in each of Statistics Canada's regional offices are responsible for the work of 3-4 senior interviewers. For the LFS, project managers and seniors are provided with performance reports each month for the interviewers they supervise. The reports include measures such as edit failure rates, nonresponse and costs. This continual feedback improves data collection procedures, thereby having a positive impact on response rates. For social surveys, there was no alternative to the dispersed organization before the advent of telephone survey methods. From 1985-1989 a program of research and testing of telephone survey methods was carried out (Drew 1991), in which a mixed organization was considered. Under this organization the role of local interviewers would be restricted largely to one of conducting face-to-face interviews, and

telephone interviewing would be carried out from the regional offices. The mixed organization would provide less opportunity for face-to-face follow-up of households that could not be contacted by telephone, leading to somewhat higher nonresponse. Also, the mixed organization would have higher overhead costs for extra office space and equipment in the regional offices. It would result in a much smaller field force, reducing the flexibility to carry out large scale *ad hoc* surveys requiring face-to-face interviews. Also, the pool of experienced field staff would be reduced to tap into each 5 years for the census of population. Based on these considerations, it was decided to retain the dispersed organization.

3.3.2 Interviewers

When new interviewers are hired for the Labour Force Survey, they are paid for 5 hours of home exercises and reading material, followed by three days of classroom training. During their first two days of interviewing in each first two months, new interviewers are observed by the senior interviewer. In addition, interviewers are routinely provided with material to read at home, and with exercises to complete dealing with different aspects of the survey taking procedures. Also, home studies are available to deal with specific problems identified in head office editing of the data. All interviewers receive an additional three days of classroom training per year. For supplements, training generally takes the form of reading material and self-study exercises to complete at home. For business surveys, the number of interviewers is much smaller, 260 in total. Training and monitoring are similar to those in the Labour Force Survey.

In a comprehensive study of nonresponse, Gower (1979) found that nonresponse rates vary greatly among interviewers. Particularly of interest, Gower found that about 15% of interviewers regularly encounter little or no nonresponse to the LFS. A focus group study is planned involving groups of superior and average interviewers. It will determine how they differ both in terms of locating respondents and in convincing them to participate in the survey. The latter will be looked at from the point of view of compliance theory, drawing on the work of Cialdinni (1991). The objective will be to identify techniques being used by superior interviewers so as to teach them to other interviewers.

3.3.3 Mode of Collection

Statistics Canada places high priority on allowing respondents to choose the mode of reporting that best fits their circumstances, including the official language of their choice. Such flexibility helps in improving response rates.

Business surveys conducted at Statistics Canada can be classified in two main groups: annual and sub-annual surveys. For the annual surveys, most of the data collection

is via questionnaire mailout and mailback administered from Ottawa, with some respondents providing data via magnetic tapes or floppy disks. The timing for mailout of annual business surveys should be linked to the respondent's fiscal year end for tax reporting purposes. This is because the required data are readily available at this time, and ambiguity about the reference year is minimized. Bilocq and Fontaine (1988), in a study on the Annual Census of Manufactures, found that the best response rates were obtained by contacting respondents three months after their fiscal year end. This implies a staggered mailout that takes fiscal year end into account. For sub-annual business surveys, data collection is mostly by mailout from Head Office and mailback to the regional offices. Most of the non-mail units respond by telephone to the regional offices, while a few respondents provide computer readable responses directly to Ottawa. It is important to respect bookkeeping practices of respondents. Most respondents use the calendar month for bookkeeping, whereas others use four and five week cycles. In both cases, data are usually available to the survey agency one or two weeks after the end of the monthly period. Telephone interviewing is used to collect data in business surveys for a variety of reasons that range from clarification of instructions to follow-up action. The quality of response may suffer if this mode of collection is improperly used. For instance, a respondent may be forced to estimate the data due to lack of availability of records near the telephone. If telephone interviewing is used on a periodic basis, such as in monthly surveys, then a best day and time arrangement with the respondent will improve response rates as well as the quality of response.

For social surveys, such as the Labour Force Survey, the mode of collection is "warm" telephone interviewing, that is, households receive an initial face-to-face interview during their first month in the sample, with predominantly telephone interviews in later months. When the initial contact with the household is made, the interviewer presents his/her identification badge. The respondent is then provided with a description of the purposes of the survey, and given assurances of the confidentiality of the responses before proceeding with the interview. The face-to-face visit is preceded by an advance letter from the Regional Director, notifying the household of its selection in the survey and describing the purpose of the survey. Respondents are invited to call on a toll free number if they have any questions before or during the survey. In a program of research and testing of telephone survey methods from 1985-1989, the feasibility of replacing the initial face-to-face interview with a telephone interview was examined. The alternative of conducting the LFS as a central telephone survey led to a 68-75% increase in nonresponse rates. There was evidence of increased nonresponse bias stemming from differences in the labour force characteristics of respondents and the additional nonrespondents (Drew 1991). The only

recurring household survey at Statistics Canada to use telephone survey methods for all its data collection is the annual General Social Survey (GSS). It uses Random Digit Dialing (RDD) in a survey of 10,000 households. On occasion the GSS sample has been augmented with households rotated out of the LFS. For example, a sample of elderly persons who had been in the LFS was selected during one round of the survey when this age group was of special interest.

3.3.4 Questionnaire Design and Introductory Material

Good questionnaire design practices contribute not only to the accuracy of the data collected, but also to the response rates. The questionnaire and introductory material are particularly important in mail collection since they are the only contact with the respondent. Material sent to respondents should include a description of the purposes of the survey, the authority under which it is conducted, assurances of confidentiality of responses, and a phone number in the agency for answering any queries on the survey questionnaire.

Questionnaires should go through a review process that is independent of the questionnaire design. This process takes the form of peer reviews by experts within the agency or focus groups of survey participants. The use of focus groups or cognitive research has resulted in several improvements aimed at respondent motivation. It has also resulted in simplification of the task of completing the questionnaires for several surveys at Statistics Canada. These include the Census of Population, the Labour Force Survey, the Census of Construction Industry and the Survey of Employment, Payrolls and Hours (Gower 1990).

3.3.5 Follow-up Strategies

For both business and social surveys, follow-up is an integral part of the overall survey design. It is only through intensive follow-up that low levels of no-contact nonresponse can be achieved. Since follow-up usually costs more per unit than primary collection (assuming a fixed survey cost), the amount of follow-up has a direct bearing on the sample size and therefore the variance, on the response rate and therefore the nonresponse bias. Design strategies range from a large sample with little follow-up to a smaller sample with intensive follow-up. In the redesign of the Monthly Wholesale and Retail Trade Survey during the 1980's, improving response rates was a priority, and this led to adoption of the strategy of a smaller sample with more intensive follow-up.

For business surveys, follow-up is undertaken both to obtain data from nonrespondents and to recontact respondents with edit failures. Most business surveys use mail as a primary mode of collection as it is inexpensive, and it gives businesses the opportunity to consult their records in responding. Nonresponse follow-up is often

restricted to a subsample of nonrespondents to reduce costs. The allocation and selection of the nonresponding units is usually based on the following factors: (i) a take-all stratum of units that must be followed-up to concentrate effort on the larger nonresponding units; (ii) an equalization of response rates across design strata; and (iii) rotation of the smaller sized nonresponding units targeted for follow-up. Nonresponse follow-up is generally by telephone for sub-annual surveys, as time constraints do not permit mail follow-up. For annual surveys, where timeliness of the collection is not as critical, mail has tended to be used for both primary collection and for initial attempts at nonresponse follow-up, with a telephone follow-up as the last resort. Increasingly, though, in recent years more of the follow-up has been by telephone for the annual surveys as well.

For social surveys, there is not as clear a distinction between primary collection and nonresponse follow-up. Follow-up consists for the most part of second and subsequent attempts to contact and interview households during the survey period. Some distinctions exist depending on the status of the dwelling. Newly sampled dwellings are initially visited to identify those that are out-of-scope and to attempt a face-to-face interview with occupants of in-scope dwellings. In cases where an interview cannot be obtained, the interviewer attempts to obtain information such as name, telephone number, and best time to call from a neighbour. Interviewers are instructed to make two to three additional attempts to interview. These follow-ups can be either by telephone or face-to-face. Occupants of previously sampled dwellings are generally interviewed by telephone. However, if repeated attempts at telephone contact are unsuccessful, a face-to-face visit is made, to insure the dwelling is still in-scope and to attempt an interview.

While follow-up is needed to bring nonresponse to acceptable levels, there is a point after which further follow-up yields diminishing returns for the money expended. There has been little work aimed at addressing the question of appropriate strategies for the scheduling and the number of follow-ups based on cost and total error considerations. Studying this issue would require cost studies to estimate parameters in a cost and mean squared error model. The factors would include contact attempts, outcomes, costs, and characteristics of respondents at different stages of follow-up. The increased automation of data collection in the years ahead should make it more feasible to collect and use such information to optimize data collection strategies.

3.3.6 Technology

Data collection for business surveys is mostly by paper and pencil. Notable exceptions are the Monthly Survey of Manufacturing where CATI is currently being used

(Coutts *et al.* 1992), and the Annual Survey of Manufacturing where CATI has been used experimentally to collect data from the smaller manufacturers. With the successful implementation of CATI for the Monthly Survey of Manufacturing, plans are under way to employ CATI for other business surveys. Experiments are also currently being carried out to test other data collection technologies for business surveys. These include: a hand-held computer for the Consumer Price Index, the Grid Pad for the Quarterly For-Hire Trucking survey, and touch tone data entry for the Survey of Employment Payrolls and Hours.

Data collection for social surveys is also based on paper and pencil technology. A decision has been taken to move to Computer Assisted Interviewing (CAI) over the next few years. The dispersed interviewing staff will be equipped with portable computers for face-to-face interviewing and for telephone interviewing from their homes. The decision was made based on positive findings from two tests of CAI on the LFS. The first test (Catlin and Ingram 1988) showed: data quality improvements such as better enumeration of persons within sampled dwellings, and fewer edit failures, with no detectable impact on survey estimates or response rates. The second test in 1991 (Coutts *et al.* 1992) demonstrated the operational viability of portable computers for CAI by interviewers in the field. Plans are to begin converting social surveys to CAI as early as 1993. These will depend on the results obtained from more extensive testing during 1992. Factors to be considered will include its impact on survey estimates, and on data quality, including response rates.

3.3.7 Response Incentives

Under the Statistics Act that sets out the legal framework governing Statistics Canada, participation in Statistics Canada surveys is mandatory for those businesses and individuals selected for survey unless the Chief Statistician designates the survey as voluntary. An example of a mandatory program is the Census of Population, where an outright refusal can lead to prosecution. For other programs, the agency relies on obtaining the co-operation of potential respondents via advance written material or publicity explaining the purpose of the survey and the confidentiality of the data, and "door step diplomacy" measures such as display of badges by face-to-face interviewers, and informing respondents about purpose and confidentiality.

Several studies of the use of response incentives have been carried out for social surveys. The first was on the Labour Force Survey (Gower 1979). In a split sample test, the Canada Handbook was given to half the households when first contacted. The result was a marginally lower refusal rate in later months for the sample receiving the incentive. Interviewers believed that the incentive was of marginal benefit, and that existing door-step procedures

were more important in reducing nonresponse. More recently, in an incentive study in the 1990 Family Expenditure Survey three treatments were administered at the interviewer level: one in which each selected household received a clipboard with the Statistics Canada logo, a second receiving the Statistics Canada publication "A Portrait of Canada," and a control sample receiving no incentive. At the national level, there was no significant change in the response rates (Kumar and Durning 1992). A study of response incentives is also planned for an upcoming longitudinal survey of income and labour.

3.4 Selective Editing

Another potential cause for nonresponse is faulty editing procedures that result in several recontacts with the respondent for the same questionnaire, lessening their willingness to cooperate on future occasions. To streamline and optimize the editing process to minimize recontacts, the following three measures should be followed. First, editing at the data capture, follow-up and imputation stages should be consistent. Second, selective editing ought to be applied to numeric data especially in business and agricultural surveys. Records that have a significant impact on the estimates are identified, and follow-up is restricted to those records. The records with a small impact should be subjected to an automated edit and imputation process to ensure consistency. Third, to keep response burden to a minimum, all errors should be identified for the units to be followed-up so that most errors can be cleared up in a single contact. The use of an inter-field edit analyzer and error localizer, such as the one in the Generalized Edit and Imputation System developed at Statistics Canada, is recommended for this requirement (Kovar, MacMillan and Whitridge 1988). If too many items fail edit but prove to be correct on follow-up, the edits should be adjusted to alleviate unnecessary response burden.

Selective editing procedures for numeric data developed at Statistics Canada can be grouped in three sets:

(i) statistical editing, (ii) grouping of variables and (iii) a score function. For statistical editing, Hidiroglou and Berthelot (1986) have developed a transformation that allows more emphasis on detecting units that show unusual changes from occasion to occasion. It recognizes that period to period changes for small units are inherently more variable than changes for large units. The cut-off bounds for edit failures are thus funnel shaped, allowing large relative changes in small units. These bounds are calculated using medians and quartiles, and are thus robust to outlier observations in the data. This method can also be used to detect outlier ratios between two variables. However, the number of pair wise comparisons can become prohibitively large. Bilocq and Berthelot (1990) recommended a method of grouping the variables into subsets of related variables and then only cross editing

variables within the subsets. The procedure used for this partitioning is based on principal component correlation methods. The significance of the errors as measured by their influence on the estimates must be considered as well. In the case of edit failure for completed questionnaires, Latouche and Berthelot (1992) have developed a score function that assigns a relative score of error importance to each respondent based on the size of the unit, the size and number of suspicious data items on the questionnaire and the relative importance of the variables. It has been demonstrated in a simulation study using this idea, that recontacting a few units is sufficient to ensure acceptable data quality for the final estimates.

3.5 Administrative Data Considerations

Response burden for Business and Agricultural Surveys at Statistics Canada is being alleviated by obtaining some data for the smaller sized units from administrative sources. Such data are also used to replace illegible, inconsistent or missing survey data. For example, the data for the smaller sized nonresponding units is imputed using tax data.

3.6 Management System for Data Collection

A good tracking system is required to determine the status of the collection process at any time. For Business Surveys, collection status codes, whose history is kept for each surveyed unit, are used to control the collection process. These collection status codes, stored in the time sequence that the survey is being carried out, are used with other codes that reflect the activity status of the unit (active, seasonal with operating dates provided, out of business, temporarily closed, *etc.*). Examples of collection status codes are: i) mode of data collection at different time points of the data collection process, ii) contact initiation codes for units (known to be active during the reference period), and for exclusions (which include closed units, out of business, temporarily closed), and iii) expected dates for return of the information to prompt additional follow-up. The management system receives information from sources external to the survey indicating a change in the status of units, and tracks the collection status from initial data collection to follow-up until all the units are ultimately classified into one of the categories under the framework described in Section 2.

4. ANALYSIS OF NONRESPONSE FOR SELECTED SURVEYS

We will briefly examine nonresponse for two surveys at Statistics Canada, to illustrate some general factors impacting on nonresponse described in Section 3.

4.1 The Monthly Retail Trade Survey

The Monthly Retail Trade Survey (MRTS) is a survey that collects sales from a sample of retail locations and inventories for a sub-sample of them. Estimates of the level and change are generated for these two variables. The sample design is a rotating simple random sample of companies stratified by province, industry and gross business income. The population size is approximately 165,000 companies, and the sample size is about 13,000. Data are collected by telephone for approximately 40% of the units and by mail for the remaining 60%. Preliminary estimates are published 7 weeks after the survey reference period, and final estimates, which include more respondents because of nonresponse follow-up are released a month later.

A redesign of the survey was implemented in January 1990. The new design differed in several aspects from the old one that had been in place since the early seventies. First, to increase the design efficiency, the number of industry groups was reduced from 34 to 18 and three size strata were used in place of two. Second, the levels of reliability were relaxed with the new design. These changes permitted a sample reduction of 35%, allowing intensive follow-up of nonrespondents. Third, data collection was decentralized to the regional offices. Under this strategy, data collection costs were higher on a per unit basis on account of the extra follow-up. There was, however, an overall gain in quality of survey results due to the reduction in nonresponse.

Both preliminary and revised weighted response rates, defined as the ratio of the estimate of sales contributed by the respondents to the estimate of sales for all in-scope units are provided in Figure 2 for the period 1986-1992. From this graph, both preliminary and revised response rates are substantially higher for the new survey than for the old survey. Preliminary rates have risen from 75% to 93%, while final rates have risen from 85% to 95%. It is also clear that the gap between the preliminary and revised response rates is much smaller for the new survey. It should be noted that in September, 1991 the preliminary rates were lower than expected because of a strike by the clerical staff handling the documents.

Several factors have contributed to the improvement in the response rates, the most important ones being mode of data collection and follow-up procedures. In the old survey, questionnaires were mailed out from and returned to Head Office (Industry Division). The mailout was carried out using manually controlled reporting arrangements. Immediate follow-up of nonrespondents was restricted to large units, and was done by telephone from Ottawa. Smaller sized nonresponding units were followed up by mail one month later, and the mail follow-up was continued for up to two additional months. Nonrespondents which had not responded for three consecutive months were referred to the regional offices for a telephone follow-up.

For the new survey, prior to their first occasion in the survey, newly sampled units (new entrants) are mailed an advance letter explaining the survey and the importance

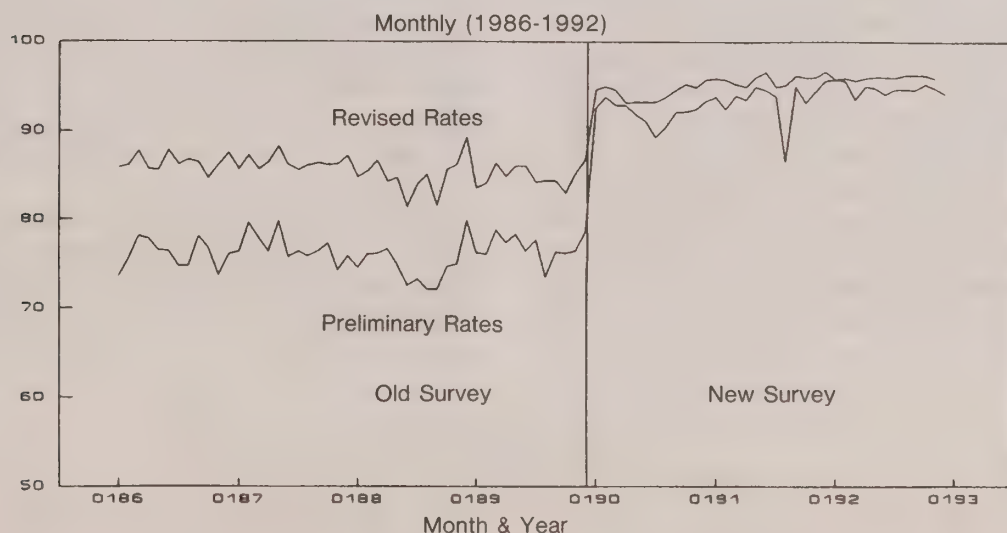


Figure 2. MRTS Response Rates

of their participation. A blank questionnaire is included. Also, each new entrant is telephoned about a week after the expected receipt of the advance letter to explain difficult ideas, to answer questions, and to offer a choice of mail or telephone data collection. For the mail respondents, questionnaires are mailed out by Industry Division using automated collection arrangements that are derived from the information on the Business Register. These collection arrangements are updated on the Business Register via profiles carried out by the Business Register Division, as well as new information found out by the regional offices during their contact with the respondent. Regional offices request data from the telephone respondents at pre-arranged dates and times, and the collected data are transmitted to Head Office after each monthly collection cycle.

4.2 Labour Force Survey

The Labour Force Survey is the largest continuous social survey conducted by Statistics Canada with a sample size of approximately 62,000 households per month. The impact of different aspects of the survey design on LFS nonresponse were discussed in Section 3. In this section, we examine historical trends in nonresponse and consider in more detail the role of nonresponse follow-up.

Table 2 below shows that the overall nonresponse rates have been steady in the 4% – 5% range throughout most of the period 1977-1991, as have refusal rates, in the 1.0% – 1.5% range. However, a few patterns are evident. One is the positive effect of the Census of Population on the nonresponse rates for the LFS, pointing to the benefit of the publicity surrounding the Census spilling over to household surveys. Nonresponse rates dropped by 1.0% between 1980 and 1981, and by 0.6% between 1985 and 1986, and by 0.4% between 1990 and 1991, the only years in which substantial drops in nonresponse rates have occurred. In 1986 virtually all the decrease was in refusals, while these accounted for over half the reduction in 1981. While the changes in nonresponse over the period are not dramatic, a gradual lessening of the positive effects of the Census is apparent. There is a slight increase in the last four years in both nonresponse and refusal rates as compared to the period from 1981 to 1987.

The graph below (Figure 3) giving the nonresponse and temporarily absent rates by month shows: the seasonal trends in the rates, with a peak in the summer months for the overall nonresponse rates, accompanied by a parallel increase in the Temporarily Absent rate. The strong relationship between the overall nonresponse rate and the Temporary Absent rate is apparent in the graph. The data collection period for the survey is normally a six day period from the Monday to the Saturday following the reference week. By Saturday of interview week the interviewers have returned all their cases to the regional offices. To reduce the seasonal peak, a Monday follow-up procedure was started in the late 1970's for the July and August surveys. Occasionally, the Monday follow-up is extended to June depending on the school year. The Monday follow-up of nonrespondents who could not be reached during the survey week is carried out from the regional offices. It has been observed that it reduces the number of cases of Temporarily Absent nonresponse.

From 1984 onwards, there has been a change in the pattern of seasonal peaks in Temporarily Absent Nonresponse. The summer peaks are less severe, but a second peak in February and March is becoming more pronounced. This seems to reflect a shift in vacation patterns of households toward more winter breaks. Consequently, in recent years, the Monday follow-up has been carried out in March if the survey week coincides with the school break.

Another noticeable feature in the LFS nonresponse pattern is higher nonresponse for households that are in the sample for the first time than for the other households. In 1980, the nonresponse rate for the first month interview households was 6.9% versus 3.5% for later months. Most of this difference occurs in the No Contact component of nonresponse. Since interviewers employ mostly face-to-face interviewing in the first month, they are limited in the number of contact attempts they can make. In later months, telephone interviewing and information obtained during the initial interview on the best time to call lead to a substantially improved contact rate.

During the 1981 post-censal redesign of the LFS, a detailed time and cost study was undertaken. The primary purpose of the study was to obtain cost information needed to carry out a cost/variance optimization of the

Table 2
LFS Nonresponse and Refusal Rates by Year

	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
NR	5.42	5.39	5.35	5.37	4.41	4.67	4.65	4.57	4.69	4.08	4.23	5.07	5.18	5.57	5.20
REF	1.34	1.45	1.41	1.47	1.16	1.19	1.14	1.18	1.18	0.99	1.06	1.30	1.31	1.51	1.38

NR = Overall Nonresponse rate.
REF = Refusal rate.

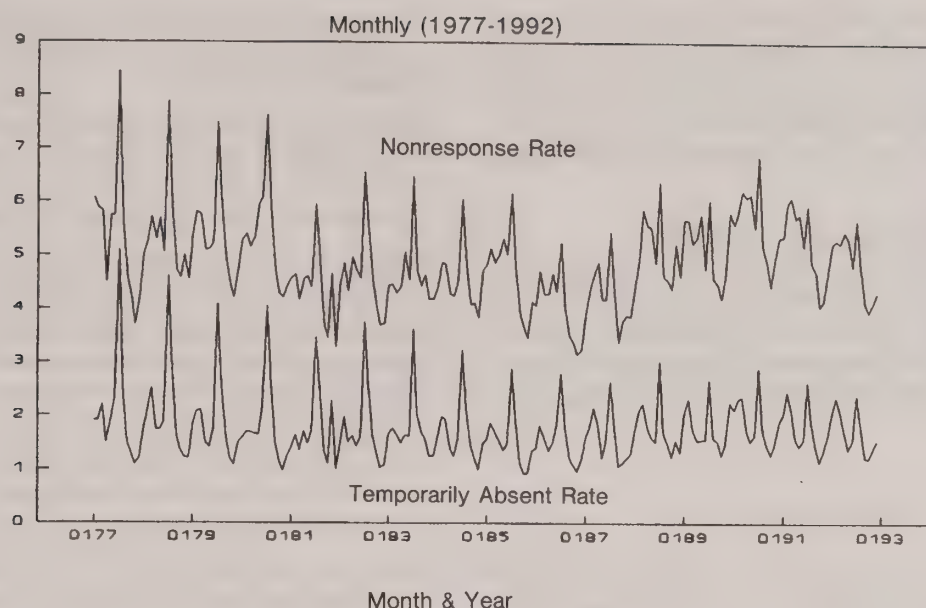


Figure 3. LFS Nonresponse Rates

survey design. The study reported by Lemaître (1983) also yielded interesting information on interviewer movement and household visit patterns, and the effect of nonresponse follow-up on response rates under face-to-face interviewing. He found a response rate of 92.4% was achieved after 3 visits, with the ratio of responses per visit consistently high at 56-61% for each round of visits. More extensive follow-up was carried out for only 3.5% of dwellings. These dwellings were visited on average another 2.5 times, with only 29% of such visits resulting in a response. The extra visits for these households accounted for 5.8% of all dwelling visits, and increased the response rate by 3.1% to 95.1%.

The 1983 Time and Cost Study was undertaken before the introduction of telephone interviewing in smaller urban and rural areas for non-first month in the sample cases, and before the introduction of telephone follow-up of first month nonresponse cases. Consideration is being given to repeating the study under the current survey conditions. One of the questions such a study could address is the cost benefit of extra visits to reduce nonresponse rates. While fourth and subsequent visits may not represent a high proportion of visits, their contribution to collection costs may be considerably higher due to the dispersion of such dwellings. Costs of such visits, coupled with information on their characteristics relative to those of other respondents, would permit an assessment of how much follow-up is warranted based on cost and mean squared error considerations.

5. SUMMARY

In this paper we have presented standards for the definition of nonresponse. In a pilot study of 7 major business and social surveys at Statistics Canada, no difficulties were found in applying the standard definitions. Beginning with the 1993 reference year, information on nonresponse for major surveys according to these standards will be reported and maintained in a central repository within the agency. This will facilitate analysis of global trends affecting response and nonresponse to surveys.

We have discussed what measures can be taken in various aspects of the survey design to help minimize nonresponse, and have illustrated their application for two major recurring surveys. Although we have restricted our focus to the role such measures play in nonresponse, they constitute good survey taking practice whose benefits encompass more than improved response rates.

In speculating about what the future holds for survey response rates in Canada, there is nothing in current trends to be alarmed about, despite a slight increase in nonresponse rates for social surveys over the last decade. However, Statistics Canada is pursuing cognitive research efforts in nonresponse aimed at better understanding respondents' attitudes and concerns about issues such as privacy, confidentiality, response burden, and record linkage. Selective editing studies are also being undertaken to focus on editing and follow-up efforts on large units. There is much scope for reducing response burden and

costs, with little impact on estimates. Findings from these studies will be helpful in designing our surveys and statistical programs in ways that respect respondents' concerns. This will permit us to continue the high levels of cooperation from the Canadian public and businesses.

ACKNOWLEDGMENTS

The authors thank B.N. Chinnappa, Statistics Canada, and the referees for their helpful comments, and to Statistics Canada: Methods and Standards Committee for its guidance and support in the development of the nonresponse framework.

REFERENCES

- BILOCQ, F., and BERTHELOT, J.-M. (1990). Analysis on Grouping of Variables and on the Detection of Questionable Units. Methodology Branch Working Paper, BSMD, 90-005E. Statistics Canada,
- BILOCQ, F., and FONTAINE, C. (1988). Étude sur la mise à la poste échelonnée pour le recensement des manufacturiers. Statistics Canada report.
- CIALDINNI R.B. (1991). Deriving Psychological Concepts relevant to survey participation from the literatures on compliance, helping and persuasion. International Workshop on Household Survey Non-response, Sweden, October 1990.
- COLLEDGE, M.J. (1989). Coverage and classification maintenance issues in economic surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, 80-107.
- CATLIN, G., and INGRAM, S. (1988). The effects of CATI on cost and quality. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*). New York: Wiley, 437-450.
- COUTTS, M., JAMIESON, R., WILLIAMS, B., and BRASLINS, A. (1992). The building of an integrated collection operation in Statistics Canada's regional offices. *Proceedings of the 1992 Annual Research Conference*. US Bureau of the Census, 395-411.
- DREW, J.D. (1991). Research and testing of telephone surveys methods at Statistics Canada. *Survey Methodology*, 17, 57-68.
- DREW, J.D., and GRAY, G.B. (1991). Standards and guidelines for definition and reporting of nonresponse to surveys. Prepared for the Second International Workshop on Household Survey Non-response, Washington, DC.
- GOWER, A.R. (1979). Nonresponse in the Canadian Labour Force Survey. *Survey Methodology*, 5, 29-58.
- GOWER, A., and ZYLSTRA, P.D. (1990). The use of qualitative methods in the design of a business survey questionnaire. Presented at the *International Conference on Measurement Errors in Surveys*, Tucson, Arizona.
- HIDIROGLOU, M.A., and BERTHELOT, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12, 73-83.
- JULIEN, C., and MARANDA F. (1990). Sample design of the 1988 National Farm Survey. *Survey Methodology*, 16, 117-129.
- KOVAR, J.G., MACMILLAN, J.H., and WHITRIDGE P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Methodology Branch Working Paper, BSMD, 88-007E. Statistics Canada.
- KUMAR, S., and DURNING, A. (1992). The Impact of Incentives on the Response Rates for FAMEX 1990: an Evaluation. Methodology Branch Working Paper, SSMD 92-001E. Statistics Canada.
- LATOUCHE, M., and BERTHELOT, J.-M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8, 389-400.
- LEMAÎTRE, G. (1983). Results from the Labour Force Survey Time and Cost Study. Internal report, Household Survey Methods Division, Statistics Canada.
- PLATEK, R., and GRAY, G.B. (1986). On the definitions of response rates. *Survey Methodology*, 12, 17-27.
- SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register for coverage improvement in the 1991 Canadian Census. *Survey Methodology*, 18, 127-141.

Double Sampling for Stratification

R.P. TREDER and J. SEDRANSK¹

ABSTRACT

Double sampling is a common alternative to simple random sampling when there are expected to be gains from using stratified sampling, but the units cannot be assigned to strata prior to sampling. It is assumed throughout that the survey objective is estimation of the finite population mean. We compare simple random sampling and three allocation methods for double sampling: (a) proportional, (b) Rao's (Rao 1973a,b) and (c) optimal. There is also an investigation of the effect on sample size selection of misspecification of an important design parameter.

KEY WORDS: Optimal sample sizes; Two phase sampling.

1. INTRODUCTION

Suppose we wish to estimate the finite population mean in a stratified population, but the units cannot be assigned to strata prior to sampling. Typically, the number of units in each stratum is unknown. Then, double sampling is commonly considered as an alternative to simple random sampling. With double sampling, a simple random sample of size n' is selected from a finite population of N units with n'_i units identified as members of stratum i , $i = 1, \dots, L$. The second phase sample is a set of L independent simple random subsamples where, in stratum i , n_i units are selected from the n'_i identified in the first phase. Letting y_{ij} denote the value of Y for the j -th unit in the second phase sample in stratum i , the finite population mean, \bar{Y} , is estimated by

$$\hat{\bar{Y}} = \sum_{i=1}^L w_i \bar{y}_i,$$

where $w_i = n'_i/n'$ and $\bar{y}_i = \sum_{j=1}^{n'_i} y_{ij}/n_i$.

Let $\sigma(n'_i)$ and $\sigma(n_i)$ denote, respectively, the set of values for first phase and second phase sample units in stratum i , $n' = (n'_1, \dots, n'_L)$ and $\sigma(n')$ the set of values for all first phase sample units. Also, let $\bar{y}_{n'}$ be the mean of the values in $\sigma(n')$, \bar{y}'_i the sample mean of $\sigma(n'_i)$, $s_i'^2 = \sum_{j=1}^{n'_i} (y_{ij} - \bar{y}'_i)^2 / (n'_i - 1)$ the sample variance of $\sigma(n'_i)$, $S_i^2 = \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2 / (N_i - 1)$ the population variance in stratum i and S^2 the analogous finite population variance. It is assumed throughout that n' is sufficiently large that $Pr(n'_i = 0)$ is negligible. Noting that $1 \leq n_i \leq n'_i$,

$$E(\hat{\bar{Y}}) = E_{\sigma(n')} \{ E(\hat{\bar{Y}} | \sigma(n')) \} = \bar{Y}$$

and

$$\begin{aligned} V(\hat{\bar{Y}}) &= V_{\sigma(n')} E\{ \hat{\bar{Y}} | \sigma(n') \} \\ &\quad + E_{\sigma(n')} \{ V(\hat{\bar{Y}} | \sigma(n')) \} \\ &= V_{\sigma(n')} (\bar{y}_{n'}) \\ &\quad + E_{\sigma(n')} \left\{ \sum_{i=1}^L w_i^2 s_i'^2 \left(\frac{1}{n_i} - \frac{1}{n'_i} \right) \right\} \end{aligned} \quad (1.1)$$

$$\begin{aligned} &= S^2 \left(\frac{1}{n'} - \frac{1}{N} \right) \\ &\quad + E_{n'} \left\{ \sum_{i=1}^L w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n'_i} \right) \right\}. \end{aligned} \quad (1.2)$$

We assume the linear cost function

$$C = c'n' + \sum_{i=1}^L c_i n_i, \quad (1.3)$$

where c' is the per unit cost associated with sampling a first phase unit, and c_i is the per unit cost of measuring Y in stratum i . The sample sizes, n' and the n_i , are selected subject to fixed total cost or to fixed total expected cost.

In this paper we compare three double sampling designs, differentiated by the way that the sample sizes, n' and the n_i , are chosen. We also compare these methods with a simple random sample having the same fixed total cost.

The alternative designs are presented in Section 2 and compared in Section 3. Section 4 presents the results of an investigation of the effect on sample size selection of misspecification of an important design parameter.

¹ R.P. Tredler, Statistical Sciences, Inc. Seattle, Washington; J. Sedransk, State University of New York at Albany, Albany, New York.

2. ALTERNATIVE METHODS

2.1 Proportional Allocation

For proportional allocation, $n_i = nw_i$ where $n = \sum_{i=1}^L n_i$. Then, using (1.2), the variance of \hat{Y} under proportional allocation, V_P , can be shown to be

$$V_P = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) \sum_{i=1}^L W_i S_i^2, \quad (2.1)$$

where $W_i = N_i/N$ is the population proportion of units in stratum i . Substituting $n_i = nw_i$ in (1.3), the expected total cost is

$$\bar{C}_P = c'n' + cn, \quad (2.2)$$

where $c = \sum_{i=1}^L W_i c_i$. Choosing n' and n to minimize (2.1) subject to fixed total expected cost, $\bar{C}_P = C^*$, yields

$$n' = \frac{C^*}{c' + \sqrt{c'cG}}, \quad (2.3a)$$

$$n = \frac{C^*}{c + \sqrt{c'c/G}}, \quad (2.3b)$$

where $G = S_W^2/S_B^2$, $S_W^2 = \sum_{i=1}^L W_i S_i^2$ and $S_B^2 = S^2 - S_W^2$.

Using (2.3),

$$V_P = \frac{1}{C^*} \left\{ \left(c' + \sqrt{c'cG} \right) S_B^2 + \left(c + \sqrt{c'c/G} \right) S_W^2 \right\} - \frac{S^2}{N}. \quad (2.4)$$

2.2 Rao's Allocation

Rao (1973a,b) proposes selecting $n_i = v_i n'_i$ where the v_i ($0 < v_i \leq 1$) are constants fixed in advance of sampling. Using this allocation in (1.2), the variance of \hat{Y} under Rao's allocation, V_R , can be shown to be

$$V_R = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \frac{1}{n'} \sum_{i=1}^L W_i S_i^2 \left(\frac{1}{v_i} - 1 \right). \quad (2.5)$$

The corresponding expected cost, \bar{C}_R , is

$$\bar{C}_R = c'n' + n' \sum_{i=1}^L c_i v_i W_i. \quad (2.6)$$

The v_i which minimize (2.5) subject to $\bar{C}_R = C^*$ satisfy

$$v_i^0 = \frac{S_i \sqrt{c'}}{S_B \sqrt{c_i}}, \quad (2.7)$$

provided that the right side of (2.7) does not exceed 1 for any i . Otherwise, an algorithm is required to determine the optimal v_i (see Rao 1973a,b). Since Rao minimizes the *unconditional* variance, the optimal v_i do not depend on the observed n'_i . After determining the v_i , n' is obtained from (2.6). Assuming that $v_i^0 \leq 1$ for each i ,

$$V_R = \frac{1}{C^*} \left(\sum_{i=1}^L W_i S_i \sqrt{c_i} + S_B \sqrt{c'} \right)^2 - \frac{S^2}{N}. \quad (2.8)$$

2.3 Optimal Allocation

The optimal allocation of the sample sizes can be obtained by minimizing (1.2) directly. For *fixed* n' and n' , select the n_i to minimize

$$\sum_{i=1}^L w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n'_i} \right), \quad (2.9)$$

subject to fixed remaining cost, $C^* - c'n' = \sum_{i=1}^L c_i n_i$ and $n_i \leq n'_i$. An algorithm is required to determine the optimal n_i given the n'_i ; see Hughes and Rao (1979) and Treder (1989). One may find the optimal value of n' by evaluating (1.2) for a sequence of "trial" values of n' . For each such n' , one estimates the expected value of (2.9) using Monte Carlo sampling of n' (see Booth and Sedransk (1969) and Treder (1989)). Note that the algorithm needed to find the optimal n_i is straightforward, and the Monte Carlo sampling of n' given n' is simple. There are several differences between the optimal allocation and Rao's allocation. In the former, total costs will not exceed C^* while in the latter the allocation only guarantees that the budget will be satisfied on the average. In the latter, the v_i are fixed in repeated sampling while in the former, allocation of the n_i depends on the observed n' . Of course, additional effort (*i.e.*, the Monte Carlo sampling) is needed to find the optimal allocation. In contrast to the optimal allocation, Rao's method permits selection of the second phase sampling fractions *prior* to observing the n'_i (see (2.7)). See Sections 3 and 4 for additional discussion.

3. COMPARISONS

3.1 Proportional vs Rao's Allocation

Assuming that $v_i^0 \leq 1$, $i = 1, \dots, L$, and using (2.4) and (2.8), it can be shown that

$$V_P - V_R = \frac{1}{C^*} \left(S_W - \frac{\bar{S}_c}{\sqrt{c}} \right) \times \left\{ 2S_B \sqrt{c'} c + c \left(S_W + \frac{\bar{S}_c}{\sqrt{c}} \right) \right\}, \quad (3.1)$$

Table 1
Percent decrease in variance, R , for Rao's allocation compared to proportional allocation for a selection of textbook examples

Reference	L	S^2	S_W^2	G	C^*	R			
						for $c' = 1$ and $c =$			
						1	2	5	25
Cochran (1977), p. 93	2	52,448	17,646	0.51	30	15.1	16.6	18.6	21.6
Hansen <i>et al.</i> (1953), p. 205	3	2,835,856	1,467,632	1.07	1,000	48.7	55.1	62.3	70.9
Sukhatme <i>et al.</i> (1984), p. 118	4	72,238	23,509	0.48	100	11.8	13.5	15.7	18.9
Cochran (1977), p. 111	7	619	343	1.25	1,000	11.2	11.7	12.4	13.7
Hansen <i>et al.</i> (1953), p. 202	8	47,393	45,595	25.36	1,000	10.5	11.0	11.5	12.0
Hansen <i>et al.</i> (1953), p. 202	11	47,393	44,974	18.59	1,000	22.9	24.1	25.4	26.7
Hansen <i>et al.</i> (1953), p. 235	11	2,039,184	820,722	0.67	1,000	21.3	24.8	29.1	35.1
Hansen <i>et al.</i> (1953), p. 202	12	47,393	40,252	5.64	1,000	16.7	18.3	19.8	21.6

Note: $R = 100(V_P - V_R)/V_P$ with V_P and V_R defined in (2.1) and (2.5) and C^* is the total budget. The cost function is defined in (1.3), and the variances (S^2 , S_W^2 , G) in (2.3).

where $\bar{S}_c = \sum_{i=1}^L W_i S_i \sqrt{c_i}$. Recalling that $c = \sum_{i=1}^L W_i c_i$ and using the Cauchy-Schwarz inequality, $S_w - \bar{S}_c/\sqrt{c} \geq 0$. Thus, as expected, $V_P - V_R \geq 0$. Defining $\bar{S} = \sum_{i=1}^L W_i S_i$ and $\bar{S}_\gamma = \sum_{i=1}^L W_i S_i \sqrt{\gamma_i}$ with $\gamma_i = c_i / \sum_{j=1}^L W_j c_j$, and using (3.1), it can be shown that

$$V_P - V_R = \frac{1}{C^*} \left\{ 2\sqrt{c'}c \left(\frac{S_B}{S_W + \bar{S}} \right) + c \right\} \times (S_W^2 - \bar{S}^2) \\ + \frac{1}{C^*} \left\{ 2\sqrt{c'}c S_B + c(\bar{S} + \bar{S}_\gamma) \right\} \times (\bar{S} - \bar{S}_\gamma). \quad (3.2)$$

The first term in (3.2) is the reduction in variance if all sampling costs are equal while the second term in (3.2) is the reduction if all strata variances are equal. As expected, if $c_i = c$ and $S_i = S$, $V_P = V_R$.

We present in Table 1, the values of $R = 100(V_P - V_R)/V_P$ corresponding to a set of textbook examples with $c_i = c$. In parallel columns we give characteristics of the associated populations (L , S^2 , S_W^2 , $G = S_W^2/S_B^2$) and C^* together with the values of R corresponding to $c/c' = 1$, 2, 5 and 25. This set of examples represents a broad range of conditions where stratified sampling may be used. For a given value of c , the range of values of R indicates the wide range of gains that may be attained. It is clear from

Table 1 that there may be substantial reductions in variance if one uses Rao's allocation, even when second phase strata sampling costs are equal and in situations when the stratification is not especially effective (note the large values of G for three examples). As c increases, R increases at a rate that is approximately constant (see Table 1).

3.2 Comparisons with Simple Random Sampling

For comparability with Rao and proportional allocations, assume a simple random sample of size n^* with expected cost $n^* \sum_{i=1}^L W_i c_i = n^*c$ (see (1.3)). Thus, for a fixed expected cost, C^* , $n^* = C^*/c$ and

$$\text{Var}(\bar{y}_{n^*}) = S^2 \left(\frac{c}{C^*} - \frac{1}{N} \right) \equiv V_S, \quad (3.3)$$

where \bar{y}_{n^*} is the sample mean. Using (2.4) and (3.3),

$$V_S - V_P = \frac{1}{C^*} \left\{ (c - c')S_B^2 - 2S_B S_W \sqrt{c'c} \right\}. \quad (3.4)$$

It can be shown that $V_S - V_P \geq 0$ if, and only if,

$$\frac{c}{c'} \geq \left(\sqrt{G} + \sqrt{1 + G} \right)^2 = LB_P, \quad (3.5)$$

where $G = S_W^2/S_B^2$. Using (2.8) and (3.3),

Table 2

Percent decrease in variance for proportional (R_P) and Rao's (R_R) allocation compared to simple random sampling for a selection of textbook examples

Reference	L	LB_P	LB_R	R_P			R_R		
				$c = 1$	5	25	$c = 1$	5	25
Cochran (1977), p. 93	2	3.8	2.6	-177.9	11.9	45.7	-136.0	28.3	57.4
Hansen <i>et al.</i> (1953), p. 205	3	6.1	1.1	-102.8	-6.1	26.4	-4.1	59.9	78.6
Sukhatme <i>et al.</i> (1984), p. 118	4	3.7	2.7	-132.8	12.8	46.6	-105.3	26.5	56.7
Hansen <i>et al.</i> (1953), p. 210	4	17.4	0.7	-127.7	-21.3	3.6	23.0	58.9	69.4
Cochran (1977), p. 111	7	6.8	4.5	-197.8	-9.8	23.3	-164.5	3.9	33.8
Hansen <i>et al.</i> (1953), p. 202	8	103.4	5.6	-38.2	-14.1	-4.0	-23.7	-0.9	8.5
Hansen <i>et al.</i> (1953), p. 202	11	76.4	1.7	-44.0	-15.6	-3.9	-11.0	13.7	23.8
Hansen <i>et al.</i> (1953), p. 235	11	4.5	2.2	-105.8	4.0	37.9	-62.0	32.0	59.7
Hansen <i>et al.</i> (1953), p. 202	12	24.5	4.0	-71.6	-19.9	0.2	-42.8	3.9	21.8

Note: Using (2.4), (2.8) and (3.3), $R_P = 100(V_S - V_P)/V_S$, $R_R = 100(V_S - V_R)/V_S$, and (LB_P, LB_R) are defined in (3.5) and (3.7). For these examples, $c' = 1$ and C^* , the total budget for each of the methods, is as in Table 1.

$$V_S - V_R = \frac{c}{C^*} \left\{ S^2 - \left(\bar{S}_\gamma + S_B \sqrt{c'/c} \right)^2 \right\}, \quad (3.6)$$

where it is again assumed that $v_i^0 \leq 1$ for all i (see (2.7)). It is easily seen that $V_S - V_R \geq 0$ if, and only if,

$$\frac{c}{c'} \geq \frac{S_B^2}{(S - \bar{S}_\gamma)^2} = LB_R. \quad (3.7)$$

In practice, one will estimate LB_P and LB_R in (3.5) and (3.7) and compare them with the cost ratio, c/c' , to decide if it will be beneficial to use double sampling with proportional or Rao's allocation rather than simple random sampling. In Table 2 we present the values of LB_P and LB_R for each of the examples in Table 1. We also include for $c = 1, 5$, and 25 the values of R , the per cent reduction in variance accruing from using a double sampling method rather than simple random sampling. As noted above, this set of examples represents a broad range of conditions where stratified sampling may be used. For a given value of c , the range of values of R_P and R_R indicates the wide range of gains (over simple random sampling) that may be obtained.

While $LB_P \geq LB_R$ is true in general, $LB_P \gg LB_R$ for many of the examples. The results point to potentially large gains for double sampling, especially using Rao's allocation, when c/c' is large. Conversely, if c/c' is relatively small, gains are modest and, in some cases, simple random sampling is preferred. This argues for careful estimation of LB_P , LB_R and c/c' .

3.3 Optimal vs Rao's Allocation

To compare the optimal allocation with that proposed by Rao, we have considered a wide range of values of the design parameters c' , S^2 and $\{(c_i, S_i^2, W_i) : i = 1, \dots, L\}$. We took $C^* = 1,000$ and considered $L = 2$ and 3. The values of the design parameters for $L = 2$ are listed in Table 3. Note that for these examples $G = S_W^2/S_B^2$ ranges from 0.01 to 10.00. We assume throughout that N is sufficiently large that S^2/N in (1.2) is negligible.

Table 3

Values of design parameters for the case of $L = 2$ strata

Parameter	Values
c'	0.125, 0.250, 0.500, 1.000
c_1	1, 4, 16
c_2	16
W_1	0.5, 0.6, 0.7, 0.8, 0.9
S^2	70.4, 128, 704
S_1^2	1, 4, 16, 64
S_2^2	64

Note: All 720 combinations of the above parameters were used. In addition, we also studied all arrangements of c' , S^2 , and S_1^2 as above together with

- (a) $c_1 = 16$; $c_2 = 1, 4, 16$ and $W_1 = 0.5, 0.6, 0.7, 0.8, 0.9$,
- (b) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 1, 4, 16$; $c_2 = 16$, and
- (c) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 16$; $c_2 = 1, 4, 16$.

To ensure comparability of the two allocations we proceeded as indicated below for each specification of the design parameters.

1. Fix a single value of n' . We used both the value of n' identified as best using (a) Rao's method and (b) the optimal allocation.
2. From each of K Monte Carlo replications ($K = 200$ or 500) we obtain $\mathbf{n}' = (n'_1, \dots, n'_L)$ and then $\mathbf{n} = (n_1, \dots, n_L)$ using the optimal allocation and $\mathbf{v} = (v_1, \dots, v_L)$ from Rao's method. For the latter we use the algorithm which makes appropriate adjustments when the right side of (2.7) exceeds 1 for one or more strata.

Since neither \mathbf{n} from the optimal method nor \mathbf{n} from Rao's method ($n_i = v_i n'_i$) are necessarily integers we round the n_i and adjust them so that for each sample the budget is satisfied (up to the approximation necessitated by having integer values of n' and \mathbf{n}). We found that if these adjustments were not made there were anomalous results where the variance of \hat{Y} using Rao's allocation was less than the corresponding variance using the optimal allocation. This occurred when the total cost associated with Rao's procedure was larger than that for the optimal procedure.

3. To obtain estimates, $\bar{V}_{(c)O}$ and $\bar{V}_{(c)R}$, of the conditional variances, $E_{n'}\{\sum_{i=1}^L w_i^2 S_i^2 (1/n_i - 1/n'_i)\}$, corresponding to the optimal and Rao's allocation, we used the average of $\sum_{i=1}^L w_i^2 S_i^2 (1/n_i - 1/n'_i)$ over the K replications. The estimates of the unconditional variance, $\text{Var}(\hat{Y})$, in (1.2) are denoted by $\bar{V}_{(u)O}$ and $\bar{V}_{(u)R}$ where $\bar{V}_{(u)R} = \bar{V}_{(c)R} + (S^2/n')$.

The precision of these estimates was assessed by estimating the standard errors and coefficients of variation of $\bar{V}_{(u)R}$ and $\bar{V}_{(c)R}$. All of the standard errors were less than 0.0022. The coefficients of variation for $\bar{V}_{(u)R}$ and $\bar{V}_{(c)R}$ were below 0.0074 and 0.023, respectively. Thus, \bar{V}_u and \bar{V}_c provide precise estimates of the unconditional and conditional variances.

We present in Table 4 estimates of the per cent increase in the average unconditional variance for Rao's allocation, $I_u = 100 (\bar{V}_{(u)R} - \bar{V}_{(u)O}) / \bar{V}_{(u)O}$, for some of the design parameters listed in Table 3. We include results only for the value of n' identified as optimal by the optimal procedure. These results are typical of those seen for the other specifications in Table 3, those that we considered for the case $L = 3$, and those which use the value of n' identified as optimal by Rao's method. It is clear from Table 4 that improvements in precision are small, ranging from none to about 4%.

We obtained somewhat similar results for the per cent increase in the *conditional* variance for Rao's allocation, $I_c = 100 (\bar{V}_{(c)R} - \bar{V}_{(c)O}) / \bar{V}_{(c)O}$, where $\bar{V}_{(c)R}$ and $\bar{V}_{(c)O}$ are obtained by estimating $E\{\sum_{i=1}^L w_i^2 S_i^2 (1/n_i - 1/n'_i)\}$ using, respectively, Rao's allocation and the optimal allocation. The results, based on 200 Monte Carlo replications

and presented using boxplots in Treder (1989, Figures 2.8.2 and C.1 – C.3), can be summarized as follows. For *all* parameter specifications, the medians of the distributions of I_c are near 0. Most of the values of I_c are small: about 95% of the parameter specifications have distributions of I_c with third quartiles less than 10%. However, occasionally, there are large values of I_c : about 15% of the parameter specifications have the maximal value of I_c larger than 20%.

Table 4

Percent increase, I_u , in the average unconditional variance $\bar{V}_{(u)}$ for Rao's allocation compared to optimal allocation for a selection of design parameters with $S^2 = 70.4$, $S_2^2 = 64$, $c_2 = 16$ and $c' = 1$

S_1^2	G	c_1		
		16	4	1
a. $(W_1, W_2) = (.9, .1)$				
64	10.000	0.0	0.4	1.4
16	0.419	0.1	0.1	0.1
4	0.166	0.1	0.1	0.4
1	0.116	0.1	0.3	0.8
b. $(W_1, W_2) = (.7, .3)$				
64	10.000	0.0	0.7	3.6
16	0.760	0.0	0.2	0.7
4	0.455	0.1	0.3	1.4
1	0.394	0.0	0.7	0.9
c. $(W_1, W_2) = (.5, .5)$				
64	10.000	0.0	1.0	4.1
16	1.316	0.0	0.4	0.9
4	0.934	0.0	0.6	1.8
1	0.858	0.0	0.2	0.0

Note: $I_u = 100 (\bar{V}_{(u)R} - \bar{V}_{(u)O}) / \bar{V}_{(u)O}$. See the note to Table 1 for definitions of the costs and variances.

These results can be explained, in part, by defining the *optimal* second phase sample size in stratum i by $n_i = \xi_i(n') \cdot n'_i$ where the dependence of n_i on the observed \mathbf{n}' is emphasized by writing $\xi_i(n')$ and $0 < \xi_i(n') \leq 1$. Then, one may find the optimal allocation by choosing the $\xi_i(n')$ to minimize (for fixed \mathbf{n}')

$$\frac{1}{n'} \sum_{i=1}^L \frac{w_i S_i^2}{\xi_i(n')}, \quad (3.8)$$

subject to $\sum_{i=1}^L c_i n'_i \cdot \xi_i(n') = C^* - c' n'$ (see 2.9).

By contrast, for the Rao allocation, for fixed n' , one selects the ν_i to minimize

$$\frac{1}{n'} \sum_{i=1}^L \frac{W_i S_i^2}{\nu_i}, \quad (3.9)$$

subject to $n' \sum_{i=1}^L c_i W_i \nu_i = C^* - c'n'$, *i.e.* fixed expected cost.

Minimizing (3.8) rather than (3.9) will yield a smaller conditional and, thus, unconditional variance. However, when n' is large, the difference between (3.8) and (3.9) will be small.

3.4 Recommendations

Given reasonable estimates of the design parameters, one should first compare the cost ratio, c/c' , with lower bounds, LB_P and LB_R , in (3.5) and (3.7) to see whether it is preferable to use double sampling rather than simple random sampling. These assessments must be done carefully because inappropriate use of double sampling may result in a *reduction* in precision. If there are good estimates of the design parameters, using Rao's allocation is preferable to proportional allocation.

Given the importance of adhering to a fixed budget we recommend the use of a modification of Rao's procedure:

Use Rao's procedure to find the "optimal" value of n' . Then, given the n'_i , use the optimal allocation procedure (*i.e.* minimize (2.9)) to find the n_i . This method guarantees that the budget will be satisfied for each sample, preserves most of the (small) gain in precision from using the optimal allocation and is easy to implement.

An alternative is to use Rao's procedure to find the "optimal" values of n' and the ν_i . Then implement an algorithm to round and modify the n_i ($n_i = \nu_i n'_i$) to ensure that the budget is satisfied for each sample. Unfortunately, it is difficult to develop the part of the algorithm needed to insure against cost overruns.

However, to avoid the large values of the proportional error in the *conditional* variance (*i.e.* I_c) that occur occasionally, one must use the *optimal* values of n' and the n_i .

Each of these methods requires knowledge of some design parameters. For Rao's allocation, the optimal ν_i require that the W_i and S_i^2 be specified. One can see from (2.9) that for the optimal allocation, the optimal n_i depend on the S_i^2 but not on the W_i . However, the optimal choice of n' requires that the W_i be specified. Alternatively, Srinath (1971) and Rao (1973a) have suggested a procedure which requires knowledge of the S_i^2 but not the W_i . Clearly, Rao's allocation requires the greatest knowledge of the design parameters and Srinath's procedure the least. Since the choice of n' is, typically, robust to misspecification of design parameters (see, *e.g.*, Sedransk 1965, Section 4.2.3), the optimal method may work well in the circumstances for which Srinath's method was designed.

4. SENSITIVITY OF ALLOCATIONS TO ESTIMATION OF DESIGN PARAMETERS

The preceding analysis assumes that the sample allocations are minimally affected by errors in the specification of the design parameters. In this section we investigate, in a simple case, the effect on $\text{Var}(\hat{Y})$ of the misspecification of an important design parameter. With proportional allocation, the choice of n' and n depends only on $G = S_W^2/S_B^2$, c' and c (see (2.3)). Estimating G by \hat{G} and substituting the resulting values of n' and n from (2.3) in (2.1),

$$\frac{V_P(\hat{Y})_{\hat{G}}}{S_W^2} = \frac{1}{C^*} \left(\frac{c' + \sqrt{c'c\hat{G}}}{G} + c + \sqrt{c'c/\hat{G}} \right) - \frac{1}{N} \left(1 + \frac{1}{G} \right), \quad (4.1)$$

where G is the correct value of S_W^2/S_B^2 and \hat{G} is used only to determine n' and n .

Table 5

Per cent increase in unconditional variance, I , for proportional allocation when G is estimated by \hat{G} .
 $C^* = 1,000$, $c' = 1$ and $c_1 = c_2 = 16$

G	\hat{G}								
	1/100	1/36	1/16	1/4	1	4	16	36	100
1/100	0.0	6.0	19.8	69.3	174.9	389.8	817.3	817.3	817.3
1/36	6.2	0.0	4.4	33.1	103.8	251.4	547.7	547.7	547.7
1/16	21.9	3.9	0.0	12.1	57.1	156.2	357.9	357.9	357.9
1/4	71.7	30.7	12.6	0.0	11.8	51.2	138.5	138.5	138.5
1	128.9	67.2	37.3	7.3	0.0	7.5	35.9	35.9	35.9
4	179.1	101.6	63.3	22.3	5.7	0.0	5.4	5.4	5.4
16	210.2	123.4	80.3	33.5	12.9	2.3	0.0	0.0	0.0
36	220.4	130.7	86.0	37.4	15.7	4.0	0.0	0.0	0.0
100	225.9	134.6	89.1	39.5	17.2	4.9	0.0	0.0	0.0

Note: I is defined in (4.3), $G = S_W^2/S_B^2$ and the cost function is given by (1.3).

The optimal value of $\text{Var}(\hat{Y})$ (i.e. when using G) in (2.4) can be expressed as

$$\frac{V_P(\hat{Y})_G}{S_W^2} = \frac{1}{C^*} \left(\frac{c'}{G} + c + 2\sqrt{c'c/G} \right) - \frac{1}{N} \left(1 + \frac{1}{G} \right). \quad (4.2)$$

If $(1/N)(1 + 1/G)$ is negligible, the per cent increase in variance due to estimating G , $I = 100\{V_P(\hat{Y})_G - V_P(\hat{Y})_{\hat{G}}\} / V_P(\hat{Y})_G$, is, from (4.1) and (4.2),

$$I = \frac{(1 - G) + \sqrt{c/c'}\{\sqrt{\hat{G}} - 2\sqrt{G} + (G/\hat{G})\}}{(1 + \sqrt{cG/c'})^2} \times 100. \quad (4.3)$$

Note that (4.3) depends only on G , \hat{G} and c/c' .

We present in Table 5 the values of I for $C^* = 1,000$, $c' = 1$, $c_1 = c_2 = 16$ and nine values of G and \hat{G} . The following conclusions are based on the results in Table 2.10.1 of Treder (1989) which includes additional values of G and \hat{G} . As long as \hat{G} is within the interval $[G/4, 4G]$, using \hat{G} to find (n', n) increases the variance by no more than 15%, typically less. If \hat{G} is in the interval $[G/2, 2G]$, the increase in variance due to misspecification is about 4% or less. As G increases, the increase in variance associated with such intervals (e.g., $[G/4, 4G]$) decreases. This happens because for large G , one has $n' = n$ and both \hat{G} and G yield the same allocation. One manifestation of this result is the array of zeros in the lower right corner of Table 5. When G is small, that is when stratification is good, the sample allocation is more sensitive to incorrect specification of G than when G is large. These findings are little influenced by the values assigned to

$c_1 = c_2$. In summary, for proportional allocation, fairly large misspecifications of the design parameter (G) lead to relatively small increases in variance.

REFERENCES

- BOOTH, G., and SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, (Vol. 1). New York: John Wiley.
- HUGHES, E., and RAO, J.N.K. (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics - Theory and Methods A*, 8(15), 1551-1574.
- RAO, J.N.K. (1973a). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- RAO, J.N.K. (1973b). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 669.
- SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.
- SRINATH, K.P. (1971). Multiphase sampling in nonresponse problems. *Journal of the American Statistical Association*, 66, 583-586.
- SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S., and ASOK, C. (1984). *Sampling Theory of Surveys with Applications*, (3rd Ed.). Ames, IA: Iowa State University Press.
- TREDER, R.P. (1989). Some problems in double sampling for stratification. Unpublished Ph.D. dissertation, University of Washington.

Stratified Telephone Survey Designs

ROBERT J. CASADY and JAMES M. LEPKOWSKI¹

ABSTRACT

Two stage random digit dialing procedures as developed by Mitofsky and elaborated by Waksberg are widely used in telephone sampling of the U.S. household population. Current alternative approaches have, relative to this procedure, coverage and cost deficiencies. These deficiencies are addressed through telephone sample designs which use listed number information to improve the cost-efficiency of random digit dialing. The telephone number frame is divided into a stratum in which listed number information is available at the 100-bank level and one for which no such information is available. The efficiencies of various sampling schemes for this stratified design are compared to simple random digit dialing and the Mitofsky-Waksberg technique. Gains in efficiency are demonstrated for nearly all such designs. Simplifying assumptions about the values of population parameters in each stratum are shown to have little overall impact on the estimated efficiency.

KEY WORDS: Random digit dialing; Optimal allocation; Coverage; Relative efficiency.

1. THE CURRENT STATUS OF TELEPHONE SURVEY DESIGNS

The two stage random digit dialing design for sampling telephone households, first proposed by Mitofsky (1970) and more fully developed by Waksberg (1978), has been widely employed in telephone surveys. The Mitofsky-Waksberg technique capitalizes on a feature of the distribution of working residential numbers (hereafter referred to as WRNs) in the U.S.: specifically, the WRNs tend to be highly clustered within banks of consecutive telephone numbers. Currently, only about twenty percent of the possible telephone numbers within the known area code, three digit prefix combinations are WRNs for the United States as a whole. However, if a bank of 100 consecutive telephone numbers can be identified that has at least one known WRN then, on average, over 50 percent of the numbers in the bank will be WRNs. A technique which can identify 100-banks containing WRNs will greatly reduce the amount of screening necessary to identify telephone numbers assigned to households.

The two-stage Mitofsky-Waksberg technique starts by obtaining a list of area code, prefix combinations for the study area (available nationally from BellCore Research; see Lepkowski 1988). A frame of telephone numbers, hereafter referred to as the BellCore Research or BCR frame, is generated by appending all 10,000 four digit suffixes (*i.e.*, 0000 to 9999) to the area code-prefix combinations. The telephone numbers in the frame are grouped into banks of 100 numbers using the area code, three digit prefix, and the first two digits of the suffix to specify each bank. For example, the area code, prefix combination 313/764 will have 100 different 100-banks: 313/764-00, 313/764-01, . . . , 313/764-99. Next, a sample of 100-banks

is selected and a single complete telephone number is generated for each selected bank by appending a two digit, randomly selected, number to the bank identifier. Each of these generated telephone numbers is dialed in the first sampling stage and the residential status of each number is determined and recorded. All 100-banks for which the randomly generated number is not a WRN are discarded. A second stage sample of WRNs is selected from all 100-banks for which the randomly generated number is a WRN. Typically an equal number of numbers, say k , are generated in each bank to start the second stage sampling process. When one of these second stage numbers is found to be non-residential, it is replaced by another randomly generated number from the same bank. This process is continued until k WRNs are identified in each bank. The result is a two stage sample based on selection of 100-banks with probabilities proportional to the number of residential numbers in the bank. This methodology has proven to be an excellent technique for identifying 100-banks with WRNs.

This technique has obvious advantages. The proportion of residential numbers within the 100-banks retained for second stage sampling is much higher than for the BCR frame in general, which results in a substantial improvement in efficiency over simple random digit dialing (RDD). It only requires that the complete set of area code, prefix combinations for the study area be known, and that the study staff have access to a random number generator for sampling telephone numbers. Finally, it also affords, in principle, complete coverage of all telephone households in the study area.

The Mitofsky-Waksberg technique also has several disadvantages. For example, not every 100-bank has the required k residential numbers so the second stage random number generation can use all 99 remaining numbers and

¹ Robert J. Casady, Bureau of Labor Statistics, U.S. Department of Labor and James M. Lepkowski, Survey Research Center, University of Michigan.

still not achieve the required k WRNs. In addition, determining the residential status of each generated number, especially at the first stage, can be difficult. For instance, in many rural areas recording equipment which notifies the caller that a number is not in service is not used. Calls to unassigned numbers are switched to a "ringing" machine. In these areas it is difficult to distinguish unassigned numbers from residential numbers where no one is at home during the study period. This difficulty is more noticeable at the end of a study period due to the need to replace non-residential numbers. Numbers generated at the end of the study period as replacements for non-residential numbers at the second stage of sampling have less time to be called. A small residual of unresolved numbers accumulates at the end of the study period, and final determination of residential status is impossible within study time constraints. Procedures for handling these unresolved numbers have been proposed (Burkheimer and Levinsohn 1988), but they often detract from the simplicity of the overall method.

Many of the difficulties with the Mitofsky-Waksberg technique can be reduced in importance through pre-screening of telephone numbers and the use of computer assisted interviewing systems. However, these difficulties are not eliminated unless departures are made from the basic simplicity and/or underlying probability sampling principles of the method (see for example Potthoff 1987 and Brick and Waksberg 1991).

Alternatively, lists of published telephone numbers have been employed as a frame. These lists of published numbers are available for the entire country from commercial firms such as Donnelley Marketing Information Systems. A straightforward selection of telephone numbers from such lists provides a very high rate of WRNs (typically at least 85%) but unfortunately does not cover households with unpublished numbers. Comparisons of telephone households with and without published numbers (see, for example, Brunner and Brunner 1971) indicates that substantial bias may result.

Lists of published numbers can be employed in a manner to provide coverage of households with unlisted numbers as well. Groves and Lepkowski (1986) describe a dual frame approach in which a sample of listed numbers is combined with a random digit dialed sample through post-stratification estimation. If coverage of the population is less important, lists of published numbers can be used to identify 100-banks with at least one listed residential number, and sampling can be restricted to such banks. Survey Sampling Inc. (1986), and previously Stock (1962) and Sudman (1973) using reverse directories, selected samples of telephone numbers from this type of 100-bank. Clearly this approach does not provide complete coverage of unlisted telephone households, but it can greatly improve sampling efficiency. In fact these "truncated frame" methods have rates of residential numbers comparable or

higher than the Mitofsky-Waksberg technique, and the troublesome replacement of non-residential numbers is not needed. Unfortunately, for many survey organizations, the coverage deficiency caused by truncating the frame is considered to be unacceptable.

The purpose of this paper is to examine stratified designs for the BCR frame as an alternative to frame truncation and Mitofsky-Waksberg designs. As an example of frame stratification, the BCR frame could be partitioned into two strata: a "high density" stratum consisting of residential numbers in 100-banks with one or more listed numbers and a "low density" stratum consisting of all the remaining numbers in the BCR frame. The "cut-off" point between high and low density strata is somewhat arbitrary; a cut-off of two or more listed numbers could reduce the chance that 100-banks are inadvertently included due to a keying error in a telephone number. Direct access to all listed numbers is not required for this stratification scheme. Counts of listed numbers, or any other indicator of the presence of listed telephone numbers in a 100-bank obtained from a reverse-directory (in metropolitan areas with such commercial services) or a commercial list for the entire country, would be sufficient. Preliminary work indicates that approximately 50% of the numbers in the high density stratum are WRNs while only about 2% of the numbers in the low density stratum are WRNs. The obvious cost difference of sampling from the two strata can be exploited through differential sample allocation. The telephone numbers in the low density stratum could be further stratified by careful examination of the characteristics of the 100-banks as determined by other data available from the BCR frame and/or the Donnelley list which may result in even further sampling efficiency.

The next section examines the question of the appropriate allocation of sample between the strata when simple random sampling is utilized within each stratum. A key feature of the stratified telephone sample approach is that it permits alternative approaches to sample selection within in the different strata. Several alternatives are presented and discussed in Section 3. Section 4 presents a study of the impact of "non-optimal" sample allocation on design efficiency. The paper concludes with a general discussion contrasting the Mitofsky-Waksberg procedure and stratified designs.

2. THE ALLOCATION PROBLEM FOR STRATIFIED TELEPHONE DESIGNS

2.1 Background

We assume that the basic sampling frame is the collection of all telephone numbers generated by appending four digit suffixes to the BCR list of area-prefix codes. Further, we assume that each household in the target population

is "linked" to one and only one telephone number in the basic sampling frame (to avoid complications of unequal probability of selection).

We also assume that we have access (possibly only indirect) to a directory based, machine readable list of telephone numbers. It should be noted that because many households choose not to list their telephone numbers in a directory, any such directory based frame will not contain all of the WRNs. Directory based lists are also by nature out of date so they will omit some numbers that are currently published WRNs while including others that are no longer WRNs.

From a survey design point of view these two frames tend to be radically different. The BCR frame includes all WRNs so it provides complete "coverage" of the households in the target population, but only about 20 percent of the telephone numbers included in the BCR frame are actually WRNs. Thus, the "hit rate" (and hence sampling efficiency) will be quite low for a simple RDD sample design utilizing the BCR frame. In contrast, a typical directory/list frame covers only about 70 percent of the target households, but the hit rate is 85 to 90 percent. In general the sampling efficiency for a simple RDD design using a directory/list frame is far better than can be attained for the BCR frame using the Mitofsky-Waksberg technique. Unfortunately, the low coverage rates associated with directory based frames preclude their use in many cases.

The basic idea of the proposed stratification approach is to utilize information from the directory based frame to partition the BCR frame into two or more strata with disparate hit rates and then allocate the sample to the strata so as to minimize cost (variance) for a specified variance (cost). Hereafter the stratum with the lowest hit rate will be referred to as the residual stratum. The truncated designs discussed earlier can be included in this general type of design if we allow the allocation of no sample to the residual stratum, and use mean squared error in place of variance.

2.2 Basic Notation

Assume that the BCR frame of telephone numbers has been partitioned into H strata based on a 100-bank attribute which can be determined from either the BCR or the directory based frame of telephone numbers. The choice of 100-banks is somewhat arbitrary; banks of from 10 to 500 consecutive numbers could be considered. For the i th stratum let

P_i = proportion of the frame included in the stratum,

h_i = proportion of the telephone numbers in the stratum that are WRNs (*i.e.* the hit rate),

w_i = average proportion of WRNs in the non-empty 100-banks (*i.e.* the average hit rate for non-empty banks),

z_i = proportion of the target population included in the stratum, and

t_i = proportion of 100-banks in the stratum that contain no WRNs.

The average hit rate for the frame is given by $\bar{h} = \sum_{i=1}^H h_i P_i$ and the proportion of empty 100-banks in the frame is given by $\bar{t} = \sum_{i=1}^H t_i P_i$.

In general only the P_i 's will be known with certainty. Data from a joint research project involving the Bureau of Labor Statistics and the University of Michigan were used to provide approximate values for the parameters h_i and w_i for the two strata in the example. Values for the remaining parameters were calculated using the algebraic relationships $t_i = 1 - (h_i/w_i)$ and $z_i = h_i P_i / \bar{h}$. The approximations for all of the frame parameters for the two stratum design are given in Table 1 below; note that for the BCR frame and $\bar{h} \cong .211$ and $\bar{t} \cong .605$. The value of \bar{h} is in close agreement with that given in Waksberg (1978) but the value of \bar{t} is somewhat smaller than the .65 provided by Groves (1977). At this time it is impossible to determine which value of \bar{t} is more accurate; in fact, the value may have changed since 1977. More recently, Tucker, Casady and Lepkowski (1992) estimated the value of \bar{t} to be .616 for 10-banks which supports the lower estimate \bar{t} of for 100-banks.

Table 1

Approximate values of the frame parameters for a two stratum design based on the BCR frame and Donnelley directory list. Stratum 1 consists of all telephone numbers in 100-banks with at least one telephone number on the Donnelley list frame; stratum 2 contains all remaining numbers

Stratum	Proportion of Frame (P_i)	Proportion of Population (z_i)	Hit Rate (h_i)	Proportion of Empty 100-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
1	.3804	.9402	.5210	.0300	.5371
2	.6196	.0598	.0204	.9584	.4900

2.3 The Basic Estimation Problem, Sample Designs and Estimators

We assume the telephone numbers in the i th stratum are labeled 1 through M_i . Let

$$d_{ij} = \begin{cases} 1 & \text{if the } j\text{th telephone number in the } i\text{th stratum is a WRN,} \\ 0 & \text{otherwise.} \end{cases}$$

The variable of interest is the household characteristic Y , and y represents the value of Y for a particular household. The population parameter to be estimated is the population mean $\mu = Y./N.$, where $N. = \sum_{i=1}^H \sum_{j=1}^{M_i} d_{ij} = \sum_{i=1}^H N_i$ and $Y. = \sum_{i=1}^H \sum_{j=1}^{M_i} d_{ij} y_{ij}$. The term N_i denotes the number of WRNs in the i th stratum and $N.$ denotes the number of WRNs in the population.

Consider two sample designs: (1) simple random sampling without replacement (*i.e.* simple RDD) from the telephone numbers in the BCR frame, denoted as design D_0 and (2) stratified simple random sampling from the BCR frame (*i.e.* independent simple RDD samples are selected from each stratum), denoted as design D_1 . Under design D_0 the standard ratio estimator for μ is given $\bar{Y}_0 = \hat{Y}_0/\hat{N}_0$ where \hat{Y}_0 and \hat{N}_0 are the usual inflation estimators for Y and N , respectively. The estimator \bar{Y}_0 is asymptotically unbiased for μ and its variance is given by $\text{var}(\bar{Y}_0) \cong \sigma^2/m\bar{h}$ where m is the sample size of telephone numbers and σ^2 is the population variance of the y 's. For the design D_1 the standard ratio estimator of μ is given by $\bar{Y}_1 = \hat{Y}_1/\hat{N}_1$ where \hat{Y}_1 and \hat{N}_1 are the standard inflation estimators for Y and N under stratified sampling. The estimator \bar{Y}_1 is also asymptotically unbiased for μ and

$$\text{var}(\bar{Y}_1) \cong \sum_{i=1}^H \frac{z_i^2 \sigma_i^2 (1 + (1 - h_i)\lambda_i)}{m_i h_i}, \quad (2.1)$$

where $\lambda_i = (\mu_i - \mu)^2/\sigma_i^2$ and m_i , μ_i , and σ_i^2 are the stratum sample sizes, means, and variances, respectively.

2.4 The Cost Model

There are costs associated both with determining the value of the indicator variable d and the value of the characteristic of interest Y . The cost function for determining the indicator variable is denoted by $C_1(\cdot)$, with

$$C_1(d) = \begin{cases} c_1 & \text{if } d = 1 \\ c_0 & \text{if } d = 0. \end{cases}$$

This model allows for the possibility that the cost of determining that a telephone number is not a WRN may be different than determining that a telephone number is a WRN. In fact the cost of determining the status of telephone numbers that are WRNs is usually less. The cost of determining the value of the characteristic Y includes only the *additional cost* of determining the value of y after the value of d has been determined. Letting $C_2(\cdot, \cdot)$ represent this additional cost, with

$$C_2(d, y) = \begin{cases} 0 & \text{if } d = 0 \\ c_2 & \text{if } d = 1. \end{cases}$$

The sum $c_1 + c_2$ represents the cost of a “productive” sample selection and c_0 represents the cost of an “unproductive” selection, then, following Waksberg (1978), $\gamma = (c_1 + c_2)/c_0$ represents the ratio of the cost of a productive selection to an unproductive selection.

The total cost for sample selection and the determination of the values of Y is a random variable for both design D_0 and D_1 . Letting $C(D_0)$ and $C(D_1)$ represent the total cost of conducting a survey under the two respective designs it is straightforward to show that

$$E[C(D_0)] = mc_0(1 + (\gamma - 1)\bar{h}) \quad (2.2)$$

and

$$E[C(D_1)] = c_0 \sum_{i=1}^H m_i(1 + (\gamma - 1)h_i). \quad (2.3)$$

2.5 Optimal Allocation for \bar{Y}_1

The stratum sample allocation that minimizes $\text{var}(\bar{Y}_1)$ for a fixed expected total cost C^* (or that minimizes $E[C(D_1)]$ for a fixed variance V^*) is specified up to a proportionality constant by

$$m_i \propto \frac{z_i \sigma_i}{\sqrt{h_i}} \left(\frac{1 + (1 - h_i)\lambda_i}{1 + (\gamma - 1)h_i} \right)^{1/2}, \quad (2.4)$$

where the proportionality constant is determined by substitution into the expected cost equation (or the variance equation, as appropriate). The proportional reduction in variance, relative to RDD sampling, under optimal allocation for fixed cost C^* (or the proportional reduction in cost under optimal allocation for fixed variance V^*) is given by

$$R(\bar{Y}_1, \bar{Y}_0) \cong 1 -$$

$$\frac{\left[\sum_{i=1}^H \frac{z_i \sigma_i}{\sqrt{h_i}} [(1 + (1 - h_i)\lambda_i)(1 + (\gamma - 1)h_i)]^{1/2} \right]^2}{\bar{h}^{-1} \sigma^2 (1 + (\gamma - 1)\bar{h})}. \quad (2.5)$$

2.6 Practical Problems Associated With Optimal Allocation

The problem of specifying the values for the parameters in the allocation equations is generic to optimal allocation schemes. For our particular case there are three basic types of parameters: frame related (z_i and h_i), cost related (γ and c_0) and those specific to the variable of interest (λ_i and σ_i^2). Currently, we have a fairly good working knowledge of the frame related parameters for the two stratum example and certain other specific stratification schemes. In Section 5, we will discuss several active research projects which should further expand our knowledge in this area.

It is clear that $\gamma \geq 1$, but the actual value can vary widely. For example, in the case of a multipurpose survey information is collected for several variables, so the costs of determining the status of a telephone number, c_0 and c_1 , are in effect amortized over the variables of interest, and γ will probably be considerably larger than unity. On the other hand, if the survey is intended to collect information on only a single variable then the value of γ is probably not much larger than two or three. Waksberg (1978) considers values of γ between 2 and 20.

Potentially the variable specific parameters pose the most serious problem. Usually our knowledge regarding the values of these parameters is limited and, in the case of multipurpose surveys, we must decide which variable(s) to use for the purposes of allocation. Fortunately, in many practical applications, two factors combine to somewhat lessen this problem. First, the allocation tends to be relatively "flat" in a neighborhood of the optimum allocation so that the reduction in variance is relatively robust with respect to allocation. Secondly, in most cases the variables of interest will not be highly related to variables of the type we are using for stratification. Therefore, with caution, we assume that $\lambda_i = 0$ and $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \dots, H$. Optimal allocation is achieved by

$$m_i \propto \frac{z_i}{\sqrt{h_i}} (1 + (\gamma - 1)h_i)^{-1/2} \quad (2.6)$$

and the proportional reduction in variance is

$$R(\bar{Y}_1, \bar{Y}_0) \cong 1 - \bar{h} \frac{\left[\sum_{i=1}^H z_i \left(\frac{1 + (\gamma - 1)h_i}{h_i} \right)^{1/2} \right]^2}{(1 + (\gamma - 1)\bar{h})} \quad (2.7)$$

In the case of the two stratum example, the allocation specified by (2.6) implies that allocation relative to the residual stratum (*i.e.* m_1/m_2) is 2.54 when $\gamma = 2$ and 1.42 when $\gamma = 10$. In the first case the projected proportional reduction in variance is $R = .283$ and in the second $R = .077$. In fact, it follows from (2.7) that as the relative cost of determining the value of the variable of interest increases, the relative benefit of optimal allocation decreases.

The Mitofsky-Waksberg sample design, denoted by D_3 , employs two stages of sample selection (*i.e.* non-empty 100-banks are selected in the first stage and WRNs are selected in the second stage). Following Waksberg (1978), we let $(k + 1)$ be the total number of WRNs selected from each sample 100-bank. The Mitofsky-Waksberg estimator, denoted by \bar{Y}_3 , is unbiased for μ , and its variance is minimized when

$$k + 1 = \max \left\{ 1, \left(\frac{(1 - \rho)\bar{t}}{(1 + (\gamma - 1)\bar{h} - \bar{t})\rho} \right)^{1/2} \right\}, \quad (2.8)$$

where ρ is intra-bank correlation. Under this "optimal" within 100-bank sample allocation the reduction in variance, relative to simple RDD, for the estimator \bar{Y}_3 is given by

$$R(\bar{Y}_3, \bar{Y}_0) \cong 1 - \frac{[(1 + (\gamma - 1)\bar{h} - \bar{t})^{1/2}(1 - \rho)^{1/2} + (\rho\bar{t})^{1/2}]^2}{1 + (\gamma - 1)\bar{h}} \quad (2.9)$$

At the national level Groves (1977) reports that $\rho \cong .05$ for economic or social statistics. Using this value of ρ , together with the values of \bar{h} and \bar{t} from the two stratum example, the projected proportional reduction in variance for the Mitofsky-Waksberg procedure is $R = .281$ when $\gamma = 2$ and $R = .060$ when $\gamma = 10$.

The two methodologies appear to produce essentially identical variance reduction for both values of the cost ratio. However, too much should not be read into this simple comparison as the projected reduction for each of the procedures is based on simplifying assumptions that will not be strictly true for any application. The only inference intended is that the two procedures appear to highly competitive under a general set of circumstances typically encountered in application.

3. ALTERNATIVE SAMPLE DESIGNS

3.1 Truncated Designs

The designs presented in the previous section produce unbiased estimates of the population mean. Incorrect assumptions regarding the various frame, cost, and population parameters only affect the efficiency of the estimators, not their expectations. Unfortunately an extremely high price is paid for the assurance of unbiasedness because sampling from the residual stratum provides information on only a small proportion of the population and at a relatively high cost. For example, suppose we are willing to settle for an estimate of the population mean exclusive of those households linked to telephone numbers in the residual stratum (*i.e.* we "truncate" the original frame by eliminating the residual stratum and select a stratified RDD sample from the remaining telephone numbers). For the two stratum example the "truncated frame" would consist only of those telephone numbers in the first stratum. The hit rate for the sample from the truncated frame would be .521, in contrast to a hit rate of .211 for the entire frame. However, only about 94% of the target population would remain in scope.

In what follows we assume that the truncated frame is simply the original BCR frame less the residual stratum which (without loss of generality) we assume to be stratum H . Accordingly, for the truncated frame $\bar{h}^* = (\bar{h} - P_H h_H) / (1 - P_H)$ is the hit rate, $\bar{t}^* = (\bar{t} - P_H t_H) / (1 - P_H)$ is

the proportion of empty 100-banks and $\mu^* = (\mu - z_H \mu_H) / (1 - z_H)$ is the population mean. Let design D_4 be stratified simple random sampling from the truncated frame, and \bar{Y}_4 the standard ratio estimator of the population mean. The estimator \bar{Y}_4 is asymptotically unbiased for μ^* , and, in general, it is biased for μ . The (asymptotic) bias is given by

$$B(\bar{Y}_4) = \mu^* - \mu = \frac{z_H(\mu - \mu_H)}{(1 - z_H)}. \quad (3.1)$$

In most practical circumstances the bias tends to zero monotonically as the proportion of the target population in the residual stratum becomes small, although, as indicated by (3.1), this is not necessarily the case. In any event, since the value of $\mu - \mu_H$ is never known, an upper limit on the proportion of the population in the residual stratum is usually the key specification to be determined when considering the use of a truncated frame. For the two stratum example approximately 6% of the target population is excluded from the sampling frame and, in almost all cases, this would not be tolerable for Federal agencies.

The equations for cost, variance, allocation, and proportional reduction in variance (or cost) are essentially the same as those presented in Section 2. In fact the only modifications required for equation (2.1) and equations (2.3) through (2.7) are to replace μ by μ^* and, for $i = 1, 2, \dots, H - 1$, replace z_i with $z_i^* = z_i / (1 - z_H)$, and replace λ_i with $\lambda_i^* = (\mu_i - \mu^*)^2 / \sigma_i^2$. Obviously all sums are only over the remaining $H - 1$ strata. For the special case where only one stratum remains after truncation the proportional reduction in variance (cost) reduces to

$$R(\bar{Y}_4, \bar{Y}_0) = 1 - \frac{\bar{h}(1 + \bar{h}^*(\gamma - 1))}{\bar{h}^*(1 + \bar{h}(\gamma - 1))}. \quad (3.2)$$

Thus for the two stratum design, the proportional reduction in variance (cost) is approximately $R = .492$ when $\gamma = 2$ and $R = .206$ when $\gamma = 10$. In both cases the reduction is substantially greater than achieved by the two methods in the previous section. However, nearly 6% of the population is not covered by the frame.

In an attempt to retain the relative efficiency of truncation while reducing the magnitude of the coverage problem, BLS and the University of Michigan are investigating several alternative stratification plans in an effort to reduce the proportion of the population in the residual stratum. One promising approach calls for the partition of the residual stratum into two or more residual strata. For example, the partitioning could create a residual stratum 3 consisting of telephone numbers in 100-banks thought to be primarily assigned to commercial establishments or not yet activated for either residential or commercial use. Residual stratum 2 will now contain all other telephone

numbers in the residual stratum from the two stratum design D_2 . Estimated frame parameters for the resulting three stratum design are given in Table 2.

Table 2

Estimated frame parameters for a proposed three stratum design based on the BCR frame and the Donnelley list frame

Stratum	Proportion of Frame (P_i)	Proportion of Population (z_i)	Hit Rate (h_i)	Proportion of Empty 100-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
1	.3804	.9402	.5210	.0300	.5371
2	.2000	.0399	.0420	.9143	.4900
3	.4196	.0199	.0100	.9796	.4900

These data were used to compute the projected proportional reduction in variance for both a three stratum design and a truncated three stratum design in which Stratum 3 is excluded. These results, together with a summary of the results for the two stratum designs and the Mitofsky-Waksberg design, are presented in Table 3 below. (Although not discussed in the text, Table 3 also includes the projected reduction in variance for a cost ratio of 20.)

Table 3

Projected proportional reduction in variance (or cost) relative to simple RDD sampling for five alternative telephone sample designs

Sample Design	Proportional Reduction in Variance or Cost			Proportion of Frame not in Scope
	$\gamma = 2$	$\gamma = 10$	$\gamma = 20$	
Two Stratum	.2829	.0766	.0320	.0000
Two Stratum (Truncated)	.4917	.2055	.1189	.0598
Mitofsky-Waksberg	.2811	.0597	.0135	.0000
Three Stratum	.3001	.0866	.0389	.0000
Three Stratum (Truncated)	.4095	.1574	.0879	.0199

The proposed partitioning strategy successfully reduces the percent of the population out of scope from nearly 6% to approximately 2%. The projected proportional reduction in variance for the truncated three stratum design is approximately $R = .410$ when $\gamma = 2$ and $R = .157$ when $\gamma = 10$. From an efficiency point of view, it occupies the middle ground between the highly efficient truncated two stratum design and unbiased designs.

Of course the issue to be faced when considering such a design is the coverage problem. The design is already subject to non-coverage of the non-telephone household population. Truncating the frame may add to any non-coverage bias already due to this source. For any particular application the risk inherent in sampling from a frame that does not include all of the target population must be weighed against the potential gain in efficiency. As expected, the standard three stratum design is slightly more efficient than the two stratum design. However, the increase in efficiency is so small that it is doubtful that the added cost of partitioning the BCR frame into an additional stratum is justified except for the purpose of truncation.

3.2 Designs Using Optimal Allocation and the Mitofsky-Waksberg Procedure

The final design to be considered is based on the stratified BCR frame. Depending on the proportion of empty 100-banks in the stratum, we use simple RDD sampling in some strata and Mitofsky-Waksberg sampling in others. The motivation for this type of design is based on the following two considerations:

- (a) Mitofsky-Waksberg sampling tends to be “administratively complex”, and if the gain in efficiency is small, simple RDD is preferred.
- (b) It follows from (2.9), applied at the stratum level, that if the proportion of empty banks in a stratum is “small” then Mitofsky-Waksberg sampling offers little, if any, increase in efficiency.

Thus, we propose to utilize simple RDD sampling in strata with a “small” proportion of empty hundred banks and Mitofsky-Waksberg sampling in the remaining strata. The criterion for determining the type of sampling to be utilized is based on equation (2.8) applied at the stratum level. Specifically, if the “optimal” total number of WRNs, as determined by equation (2.8), to be selected from sample 100-banks in a particular stratum is equal to one, then the stratum is designated a simple RDD stratum; otherwise it is designated a Mitofsky-Waksberg stratum. In terms of the proportion of empty hundred banks, the i th stratum will be an RDD stratum if

$$t_i \leq \frac{2.25\rho(1 + h_i(\gamma - 1))}{(1 + 1.25\rho)} \quad (3.3)$$

and a Mitofsky-Waksberg stratum otherwise. For the two stratum example, the first stratum is a RDD stratum, and the second is a Mitofsky-Waksberg stratum for γ equal either 2 or 10.

Formally the proposed sample design is as follows. The BCR frame has been partitioned into H strata and, according to the criteria given in (3.3), simple RDD sampling is specified for the first H_1 strata and Mitofsky-Waksberg sampling is specified for the remaining strata.

Let:

- m_i = the number of telephone numbers selected from the i th RDD stratum,
- m'_i = the number of WRNs in the sample from the i th RDD stratum,
- \tilde{m}_i = the number of 100-banks selected from the i th Mitofsky-Waksberg stratum,
- \tilde{m}'_i = the number of retained 100-banks in the i th Mitofsky-Waksberg stratum,
- k_i = number of additional WRNs selected from each retained 100-bank, and
- y_i = aggregate of y values for the sample WRNs from the i th stratum.

The combined ratio estimator $\bar{Y}_5 = \hat{Y}_5 / \hat{N}_5$, where $\hat{Y}_5 = \sum_{i=1}^{H_1} M_i / m_i y_i + \sum_{i=H_1+1}^H M_i / \tilde{m}_i (y_i / k_i + 1)$ and $\hat{N}_5 = \sum_{i=1}^{H_1} M_i / m_i m'_i + \sum_{i=H_1+1}^H M_i / \tilde{m}_i \tilde{m}'_i$, is utilized to estimate the population mean μ and the values of m_i , \tilde{m}_i and k_i are to be chosen to minimize $\text{var}(\bar{Y}_5)$ or the expected cost as specified.

The estimator \bar{Y}_5 is asymptotically unbiased for μ and it is straightforward to show that

$$\begin{aligned} \text{var}(\bar{Y}_5) \cong & \sum_{i=1}^{H_1} \frac{z_i^2 \sigma_i^2}{m_i h_i} (1 + (1 - h_i) \lambda_i) \\ & + \sum_{i=H_1+1}^H \frac{z_i^2 \sigma_i^2}{\tilde{m}_i h_i} [1 + (1 - h_i) \lambda_i \\ & - k_i(1 - \rho)(k_i + 1)^{-1}] \end{aligned} \quad (3.4)$$

and

$$\begin{aligned} E[C(D_5)] = c_0 \left\{ \sum_{i=1}^{H_1} m_i [1 + h_i(\gamma - 1)] \right. \\ \left. + \sum_{i=H_1+1}^H \tilde{m}_i [1 + k_i(1 - t_i) \right. \\ \left. + h_i(k_i + 1)(\gamma - 1)] \right\}. \end{aligned} \quad (3.5)$$

The optimal values of m_i and \tilde{m}_i , specified up to a proportionality constant, are given by

$$m_i \propto z_i \sigma_i \left(\frac{1 + (1 - h_i) \lambda_i}{h_i(1 + h_i(\gamma - 1))} \right)^{1/2}, \quad (3.6)$$

for $i = 1, \dots, H_1$ and

$$\tilde{m}_i \propto z_i \sigma_i \left(\frac{\lambda_i(1 - h_i) + \rho}{h_i t_i} \right)^{1/2}, \quad (3.7)$$

for $i = H_1 + 1, \dots, H$. The optimal value of $(k_i + 1)$, for $i = H_1 + 1, \dots, H$, is given by

$$k_i + 1 = \max \left\{ 1, \left(\frac{t_i(1 - \rho)}{(1 + h_i(\gamma - 1) - t_i)(\lambda_i(1 - h_i) + \rho)} \right)^{1/2} \right\}. \quad (3.8)$$

The proportionality constant for (3.6) and (3.7) is found by substitution into the expected cost equation or the variance equation as appropriate.

Under optimal allocation the reduction in variance (or cost) relative to simple RDD, is given by

$$R(\bar{Y}_S, \bar{Y}_0) = 1 - \frac{\bar{h}\Phi^2}{\sigma^2(1 + (\gamma - 1)\bar{h})}, \quad (3.9)$$

where

$$\begin{aligned} \Phi = & \sum_{i=1}^{H_1} \frac{z_i \sigma_i}{h_i^{1/2}} (1 + (1 - h_i)\lambda_i)^{1/2} (1 + (\gamma - 1)h_i)^{1/2} \\ & + \sum_{i=H_1+1}^H \frac{z_i \sigma_i}{h_i^{1/2}} \left[(\rho + (1 - h_i)\lambda_i)^{1/2} t_i^{1/2} + \right. \\ & \left. (1 - t_i + (\gamma - 1)h_i)^{1/2} (1 - \rho)^{1/2} \right]. \quad (3.10) \end{aligned}$$

Under the simplifying assumptions $\lambda_i = 0$ and $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \dots, H$,

$$\begin{aligned} \Phi = & \sigma \left[\sum_{i=1}^{H_1} \frac{z_i}{h_i^{1/2}} (1 + (\gamma - 1)h_i)^{1/2} \right] \\ & + \sigma \left[\sum_{i=H_1+1}^H \frac{z_i}{h_i^{1/2}} ((\rho t_i)^{1/2} + \right. \\ & \left. ((1 - t_i + (\gamma - 1)h_i)(1 - \rho))^{1/2}) \right]. \quad (3.11) \end{aligned}$$

When applied to the two stratum frame, this combined sampling strategy yields a proportional reduction in variance of approximately $R = .440$ for $\gamma = 2$ and $R = .157$ for $\gamma = 10$. For both of the cost ratios, the reduction in variance is considerably larger than achieved by any of the unbiased procedures considered previously. In fact, the variance reduction is essentially equivalent to that attained by the three stratum truncated design (which is subject to a bias of unknown magnitude). Thus, on first consideration, this combined sampling strategy appears to be superior to all of the other methods.

Unfortunately there are practical problems which may preclude the use of this sampling design in certain situations. For example, the hit rate in the Mitofsky-Waksberg stratum is very low (only .02) so the number of first stage sample 100-banks must be fairly large in order that the expected number of retained 100-banks is not too small. On the other hand, the *relative* number of first stage sample units allocated to the RDD stratum is considerably larger than allocated to the Mitofsky-Waksberg stratum, therefore a large overall sample size is required (see Table 4). Also, from Table 4, the number of WRNs required from each of the retained 100-banks is relatively large and may actually exceed the number of WRNs in some banks. Clearly both of these problems are more acute for $\gamma = 2$ than for $\gamma = 10$. Therefore, the use of this design is restricted to situations where resources can support a "large" sample, and the cost ratio is moderate to large.

Table 4

First stage allocation ratios and second stage sample sizes for the combined RDD/Mitofsky-Waksberg sample design applied to the two stratum BCR frame

Stratum	$\gamma = 2$		$\gamma = 10$	
	m_1/\tilde{m}_2	Sample Size Second Stage	m_1/\tilde{m}_2	Sample Size Second Stage
1	28.17	N.A.	14.56	N.A.
2	N.A.	17.00	N.A.	9.00

4. SAMPLE ALLOCATION AND DESIGN EFFICIENCY

In Section 2.6 the problem of specifying the parameters required to optimally allocate the sample to the various strata was considered. It was noted that the variable specific parameters (*i.e.* the λ_i and σ_i^2) tend to pose the most serious problem since we usually have little information regarding their values. For most cases the variables of analytic interest will not be very highly related to the variables used for stratification. Thus it is reasonable to assume that $\lambda_i = 0$ and $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \dots, H$. Under these assumptions the optimal allocation is given by (2.6) and the proportional reduction in variance is given by (2.7).

It is obvious that for any particular application these assumptions will never be strictly true, so when we allocate according to (2.6) the actual proportional reduction in variance will not be that given exactly by (2.7). Furthermore, allocating according to (2.6) will not provide the maximum reduction in variance which is achieved under the optimal allocation specified by (2.4). Assuming that we plan to allocate according to (2.6) two questions need to be addressed: (1) does (2.7) give a reasonable approximation to the actual reduction in variance, and (2) is the actual

reduction in variance reasonably close to the maximum possible reduction in variance? A single simple answer is not possible for either question because the outcome depends on exactly how and to what extent the assumptions failed. In the following we address these question for the two stratum design under three specific cases of model failure which are typical of situations encountered in the "real world". In all three cases the results indicate strongly affirmative answers for both questions.

In the first case we assume that $\sigma_1^2 = \sigma_2^2 \equiv W^2$ but $\lambda_1 \neq \lambda_2$. The projected, the actual, and the maximum reduction in variance were computed for selected values of $\beta = |\sqrt{\lambda_1} - \sqrt{\lambda_2}| = |\mu_1 - \mu_2|/W$ between 0.00 to 0.50 and the results are presented in Table 5 below. Based on our previous discussion regarding the weak relationship between the analytic and stratification variables it would seem highly unlikely that β will ever be larger than 0.50. The results in Table 5 indicate that for both cost ratios and for all selected values of β the actual reduction in variance achieved by allocation under the simplifying assumptions is essentially equivalent to that which would be attained under "optimal" allocation. For both cost ratios the projected reduction in variance is always larger than the reduction actually attained and the difference increases as β becomes larger. However, it should be noted that for $\beta \leq .35$ the percentage difference between the projected reduction and the actual reduction is less than 10% when $\gamma = 10$, and less than 4% when $\gamma = 2$.

Table 5

The projected, the actual, and the maximum proportional reduction in variance for cost ratios of 2 and 10 and values of β between 0.00 and 0.50

β	$\gamma = 2$			$\gamma = 10$		
	Projected Reduction	Actual Reduction	Maximum Reduction	Projected Reduction	Actual Reduction	Maximum Reduction
0.00	.2829	.2829	.2829	.0766	.0766	.0766
0.10	.2829	.2820	.2820	.0766	.0761	.0761
0.20	.2829	.2793	.2794	.0766	.0745	.0746
0.30	.2829	.2748	.2750	.0766	.0720	.0721
0.40	.2829	.2686	.2692	.0766	.0684	.0689
0.50	.2829	.2607	.2619	.0766	.0639	.0649

The second general case considered assumes that the analytic variable is Bernoulli, where p_1 and p_2 represent the proportion of the population with the attribute of interest in stratum 1 and stratum 2, respectively. The projected, the actual, and the maximum proportional reduction in variance were computed for two specific cases of assumption failure, namely $p_2 = .90p_1$ and $p_2 = 1.10p_1$; p_1 was allowed to vary from .05 to .50 and cost ratios of 2 and 10 were considered.

As discussed before it is probably reasonable to assume that p_2 will be within 10% of p_1 in most "real world" situations so these results can be considered general for Bernoulli type analytic variables. The actual reduction in variance was virtually identical to that attained under optimal allocation in all cases; thus, allocation under (2.6) can be considered (near) optimal. The projected reduction in variance was also very close to the actual reduction. When p_2 was smaller than p_1 the actual reduction was always larger than the predicted reduction, and the converse was true when p_2 was larger than p_1 . In both cases the maximum difference (which was only about 3.5% of the actual reduction when $\gamma = 2$ and 8.3% of the actual reduction when $\gamma = 10$) occurred when $p_1 = 0.05$ and monotonically decreased as p_1 increased.

In summary; the two cases considered seem to indicate that so long as the assumptions which yield the allocation specified by (2.6) are not radically violated, the variance will be very near that attained under optimal allocation. Furthermore, the proportional reduction in variance given by (2.7) provides an approximation for the actual reduction in variance which is at least accurate enough for the purposes of survey design.

5. CONCLUDING REMARKS

The strengths of the Mitofsky-Waksberg technique for generating telephone samples are clear: high hit rates in the second stage of selection, an efficient method for screening empty banks of telephone numbers, and a conceptually ingenious approach to sample generation. It is a remarkable testimony to the strength of the technique that it is widely considered to be the standard method of random digit dialing with few serious competitors after many years. The weakness of the technique (first stage screening and replacement of non-residential numbers during the data collection) does not, on the surface, seem to be important relative to its general strength. However, these features can cause substantial difficulty, especially in short time-period telephone survey operations.

In this paper stratified designs, based on commercial lists of telephone numbers, are proposed as alternatives to the Mitofsky-Waksberg technique. Both two and three stratum designs are studied in detail. In addition to simple random sampling within each stratum, two general alternatives are considered:

- (1) Simple random sampling from all strata except the low density stratum frame where the Mitofsky-Waksberg method is used.
- (2) Simple random sampling from all strata except the low density stratum which is not sampled at all.

The basic thesis of this paper is that stratified sampling methods, using strata based on counts of listed telephone

numbers, are at least as efficient as the Mitofsky-Waksberg technique. Furthermore, these designs can eliminate the need for the troublesome replacement of non-residential numbers at the second stage, since the only telephone numbers that must be dialed in the high density stratum are those that are generated at the beginning of the study. Specific conclusions include the following:

- For low cost ratios, the two and three stratum designs are as efficient as the Mitofsky-Waksberg approach.
- When numbers can be dropped from the low density stratum, these alternative designs are much more efficient, but at the price of unknown bias due to excluding part of the target population.
- When cost ratios are high, the two and three stratum approaches are clearly superior.

A critical issue is the magnitude of the bias introduced by dropping the low density stratum. As noted previously, approximately 7% of U.S. households do not have a telephone and truncating the frame may add to the non-coverage bias. As less than 5% of the U.S. household population is expected to be contained in the low density stratum it is likely that the additional coverage bias will not be substantial for many characteristics of the total population. On the other hand, for some characteristics, and for some subgroups of the population, the magnitude of the additional bias may be large enough to be of concern. Further empirical investigations of this population must be conducted.

There are two costs associated with the use of stratified designs that may detract from their use: the cost of the commercial list used to stratify the BCR frame and the overall lower hit rate. The cost of stratifying the frame into high and low density strata is not addressed in this investigation because the requisite information was derived from a specialized research file. The cost of stratification is a fixed cost and therefore will reduce the resources available for data collection. It is not known what the fixed cost will be in the future as arrangements are made with commercial vendors to routinely provide such data. Furthermore, this fixed stratification cost can be amortized over multiple studies to greatly reduce its impact on any single sample. It is unlikely that data collection for one time surveys will find either the Mitofsky-Waksberg or the stratification method described here to be as cost-effective as indicated. Further investigation is needed into the frame costs before a complete answer can be found.

The second cost issue concerns the lower hit rates presented in this paper. Given the relative competitive efficiencies of the alternatives considered here, it appears that the lower hit rates do not seriously detract from the efficiency of the alternatives. It may be possible to improve the hit rates in the high density stratum if smaller banks of numbers are used. For example, in another investigation

we have found that 10-banks will have hit rates in the neighborhood of .57 compared to the .52 reported here for 100-banks. Of course, working with 10-banks substantially increases the size of files and processing operations that must be used to generate samples and the cost of a 10-bank frame is likely to be much higher than the 100-bank frame.

The cost models as shown in (2.2) and (2.3) are relatively simple, ignoring many cost differences in the telephone survey process that may be important for comparisons of relative efficiencies of the designs. These cost models allow the allocations to be expressed in a straightforward way, but they do not specifically address the cost components associated with two features of the Mitofsky-Waksberg technique that the alternative designs address; replacement of nonworking numbers and weighting to compensate for exhausted clusters. Thus, the cost models ignore structural cost differences between the Mitofsky-Waksberg approach and the proposed alternatives that, if properly taken into account, could effect the relative efficiency of the two methods.

Clearly the results presented here are insufficient to draw final conclusions about the overall value of these alternative designs. Further cost data and empirical evidence on the size of the bias caused by eliminating the numbers from the low density stratum is required before a final conclusion can be reached.

ACKNOWLEDGMENTS

The support and assistance of Clyde Tucker and Bob Groves is gratefully acknowledged. The findings and opinions expressed in this article are those of the authors and do not necessarily reflect those of the U. S. Bureau of Labor Statistics or the University of Michigan.

REFERENCES

- BRICK, J.M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, 17, 27-41.
- BRUNNER, J.A., and BRUNNER, G.A. (1971). Are voluntarily unlisted telephone subscribers really different? *Journal of Marketing Research*, 8, 121-124.
- BURKHEIMER, G.J., and LEVINSOHN, J.R. (1988). Implementing the Mitofsky-Waksberg sampling design with accelerated sequential replacement. In *Telephone Survey Methodology*, (Eds. R. Groves, et al.) 99-112. New York: John Wiley and Sons.
- GROVES, R.M. (1977). An Empirical Comparison of Two Telephone Designs. Unpublished report of the Survey Research Center of the University of Michigan, Ann Arbor, MI.

- GROVES, R.M., and LEPKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 340-345.
- LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. In *Telephone Survey Methodology*, (Eds. R. Groves, *et al.*) 73-98. New York: John Wiley and Sons.
- MITOFSKY, W. (1970). Sampling of telephone households. Unpublished CBS News memorandum, 1970.
- POTTHOFF, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.
- STOCK, J.S. (1962). How to improve samples based on telephone listings. *Journal of Advertising Research*, 2, 55-51.
- SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.
- SURVEY SAMPLING, INC. (1986). Statistical characteristics of random digit telephone samples produced by Survey Sampling, Inc. Westport, CT: Survey Sampling, Inc.
- TUCKER, C., CASADY, R.J., and LEPKOWSKI, J.M. (1992). Sample allocation for stratified telephone sample designs. To appear, *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Poisson-Poisson and Binomial-Poisson Sampling in Forestry

Z. OUYANG, H.T. SCHREUDER, T. MAX and M. WILLIAMS¹

ABSTRACT

Binomial-Poisson and Poisson-Poisson sampling are introduced for use in forest sampling. Several estimators of the population total are discussed for these designs. Simulation comparisons of the properties of the estimators were made for three small forestry populations. A modification of the standard estimator used for Poisson sampling and a new estimator, called a modified Srivastava estimator, appear to be most efficient. The latter is unfortunately badly biased for all 3 populations.

KEY WORDS: High value timber; Volume estimation; Estimators for Poisson-Poisson sampling; Simulation comparisons; Forest sampling; Srivastava estimation.

1. INTRODUCTION

Volume estimation in forestry has been highly developed in the sense that very efficient sampling strategies are available to estimate total volume (Schreuder and Ouyang 1992). Estimating and measuring defect is often not built into these strategies since measuring defect is difficult and not economically justified in most stands. But in high value stands two-phase strategies such as Poisson-Poisson sampling may be suitable where defect is measured on trees at the second phase. To sample truck loads of logs, binomial-Poisson sampling may be a suitable sampling design.

The purpose of this article is to present the theory of binomial-Poisson and Poisson-Poisson sampling and discuss some of the properties of estimators for these designs based on simulation.

2. REVIEW OF LITERATURE

Singh and Singh (1965) developed the theory for two-phase sampling with probability proportional to size (pps) sampling at the second phase. Furthermore, Särndal and Swensson (1987) gave a general theory of two-phase sampling. A list of sampling units is assumed to be available at the first phase prior to sampling.

Hajek (1957) developed Poisson sampling and Grosenbaugh (1964) suggested its use for one-phase unequal probability sampling when no list is available. Poisson sampling is a scheme such that each unit in a population, say unit i , is drawn into the sample independently with probability p_i . Thus the inclusion probability of unit i is

equal to p_i , and joint inclusion probability of units i and j is equal to $p_i p_j$. Binomial sampling, also often called Bernoulli sampling, is a special case of Poisson sampling when all p_i are equal.

In forest survey, Poisson sampling is often implemented as follows (Schreuder *et al.* 1968).

1. Visit the N units (say trees) in the population in any order and measure or ocularly estimate the value of a covariate x_i ($i = 1, \dots, N$) highly correlated with the value of interest y_i ($i = 1, \dots, N$).
2. As each x_i is observed, compare it with a random integer, δ_i , randomly selected from the range $1 \leq \delta_i \leq L$, where L is an integer selected prior to sampling. L is picked such that $L = X/n_e$ where X = total for the covariate in the population and n_e is the desired sample size. X is usually not known before sampling and needs to be estimated.
3. If $\delta_i \leq x_i$, select the unit for the sample and measure y_i .

Implementation of this method results in a sample of size n , where $E(n) = n_e$ (if a good estimate of X was made prior to sampling). In binomial sampling all the x_i ($i = 1, \dots, N$) are the same (Goodman 1949).

3. SAMPLING METHODS

The United States Forest Service Region 6 (Wendall L. Jones – personal communication) uses a truck load sampling method as follows: as trucks pull up to the mill a binomial sampling technique is used to randomly select

¹ Z. Ouyang, Formerly post-doctoral fellow, Statistics Dept., Colorado State University Fort Collins, Colorado, now Research Statistician, ICI Seeds, Inc., Slater, Iowa; H.T. Schreuder, Project Leader, Multiresource Inventory Techniques Project, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado; T. Max, Station Biometrician, USDA Forest Service, Pacific Northwest Experiment Station, Portland, Oregon; M. Williams, Statistician, Multiresource Inventory Techniques Project, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado.

trucks to be sampled, with $p = 0.10$ say. These truck loads are measured for volume. A problem with this approach is that there are long runs of no trucks being sampled. As communicated to one of the authors, this was considered highly undesirable from a practical point of view. An alternative approach, which should decrease the frequency of long runs of no samples, and could be more efficient is to use binomial – Poisson sampling instead as follows:

Apply binomial sampling with a larger p (say $p = 0.30$). The scaler visually estimates volume on the selected loads. A Poisson subsample of these loads is then selected with probability proportional to the estimated volumes and the loads selected at this phase are scaled for volume. This is binomial-Poisson sampling.

For high-value timber stands in the Pacific Northwest of the United States highly accurate estimates of net volume, that is, usable volume is often desired. Actually cutting down and destructively measuring sample trees is the most reliable method of determining net volume, *i.e.* total volume minus defective volume (Johnson and Hartman 1972). Poisson-Poisson sampling may be a good sampling design in this situation. The procedure is:

1. Select n_1 out of the N trees in the population by Poisson sampling, selecting the trees proportional to some estimate of gross volume, say $x_1 =$ diameter at breast height squared (d^2). With Poisson sampling actual sample size is random, say n_1 where $E(n_1) = n_{e1}$. Ocularly estimate say $x_2 =$ ocular net volume.
2. Select n_2 out of the n_1 sample trees proportional to x_2 , by Poisson sampling. Here $E(n_2) = n_{e2}$ is the expected sample size at the second phase.

The n_2 sample trees are then cut and destructively measured for gross, net, and defective volume. To maintain maximum efficiency in both inventory and operations it is probably best to implement both sampling phases at once and mark the n_2 sample trees at inventory time. Ascertaining usable volume for these n_2 trees is done later either by a different crew or by carrying the sample trees into a sawmill to process them for actual wood products. Binomial-Poisson sampling is a special case of this. (If the second phase is implemented separately from the first phase then a list of sampling units is available to implement the second phase and some *pps* procedure with fixed sampling size should be used instead of Poisson sampling. This approach is usually inefficient because it requires two trips to the field location).

4. NOTATION

- N = Population size (not known until sampling is completed).
 n_e = Expected sample size in one-phase Poisson sampling.

- n = Achieved sample size in one-phase Poisson sampling.
 n_{e1} = Expected sample size of first phase in two-phase Poisson sampling.
 n_1 = Achieved sample size of first phase in two-phase Poisson sampling.
 n_{e2} = Expected sample size of second phase in two-phase Poisson sampling.
 n_2 = Achieved sample size of second phase in two-phase Poisson sampling.
 Y = Total usable volume in the population (to be estimated by two-phase sampling), $Y = \sum_{i=1}^N y_i$.
 x_{1i} = Covariate value for tree i at phase 1, say tree diameter at breast height squared (D^2).
 X_1 = $\sum_{i=1}^N x_{1i}$ (known after implementing the first phase in the entire population).
 $\pi_i(P)$ = Probability of selecting tree i in one-phase Poisson sampling ($= n_e x_{1i} / X_1$). If all the $\pi_i(P)$ are equal, this is one-phase binomial sampling.
 π_{1i} = Probability of selecting tree i at phase 1 ($= n_{e1} x_{1i} / X_1$).
 x_{2i} = Covariate value for tree i at phase 2, say ocular estimate of net volume.
 X_2 = Total amount of ocularly-estimated volume in the population (only obtained for the n_1 sample trees at the first phase so X_2 can only be estimated).
 π_{2i} = Probability of selecting tree i at the second phase ($= n_{e2} x_{2i} / \sum_{i=1}^{n_1} x_{2i}$).
 y_i = Value of interest for tree i (say net volume).
 π_i = Probability of selecting tree i through both sampling phases ($= \pi_{1i} \pi_{2i}$).
 π_i^* = Approximate probability of selecting tree i through both sampling phases ($= \pi_{1i}^* \pi_{2i}^*$ where $\pi_{1i}^* = n_1 x_{1i} / X_1$ and $\pi_{2i}^* = n_2 x_{2i} / \sum_{i=1}^{n_1} x_{2i}$).

5. THEORY

For Poisson sampling, the estimator

$$\hat{Y}_u = \sum_{i=1}^n y_i / \pi_i(P), \quad (1)$$

is unbiased but very inefficient and should be replaced by the following approximately unbiased estimator (Grosenbaugh 1964):

$$\hat{Y}_a = \begin{cases} \frac{n_e}{n} \hat{Y}_u & \text{if } n > 0 \\ 0 & \text{if } n = 0. \end{cases} \quad (2)$$

The variance of \hat{Y}_a , as given in Brewer and Hanif (1983), is

$$V(\hat{Y}_a) = \sum_{i=1}^W \pi_i(P) [1 - \pi_i(P)] \left[\frac{y_i}{\pi_i(P)} - \frac{Y}{n_e} \right]^2 + p_0 Y^2,$$

where $p_0 = P(n = 0)$.

For Poisson-Poisson (PP) sampling, an estimator for Y analogous to \hat{Y}_u above is the unbiased estimator

$$\hat{Y}_1 = \sum_{i=1}^{n_2} y_i / \pi_i. \quad (3)$$

This estimator can be horribly inefficient as pointed out for \hat{Y}_u in Poisson sampling (Schreuder *et al.* 1968).

The variance of \hat{Y}_1 can be written down by using the general formulas developed by Särndal and Swensson (1987) for unbiased estimation in double sampling:

$$V(\hat{Y}_1) = \sum_{i=1}^N \left(\frac{1 - \pi_{1i}}{\pi_{1i}} \right) y_i^2 + E_1 \left\{ \sum_{i=1}^{n_1} \left(\frac{1 - \pi_{2i}}{\pi_{2i}} \right) \left(\frac{y_i}{\pi_{1i}} \right)^2 \right\},$$

where E_1 denotes expectation over the first-phase sample. Since \hat{Y}_1 is not efficient we do not give its variance estimator. Analogous to the more efficient adjusted estimator in Poisson sampling we have the approximately unbiased estimator

$$\hat{Y}_2 = \sum_{i=1}^{n_2} y_i / \pi_i^* = \hat{Y}_1 (n_{e1} / n_1) (n_{e2} / n_2). \quad (4)$$

The variance of \hat{Y}_2 is:

$$V(\hat{Y}_2) = p(\phi) Y^2 + \sum_{i=1}^N \pi_{1i} (1 - \pi_{1i}) \left(\frac{y_i}{\pi_{1i}} - \frac{Y}{n_{e1}} \right)^2 + \sum_{s_1 \neq \phi} p_1(s_1) \left\{ \sum_{i \in s_1} \pi_{2i} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{1i} \pi_{2i}} - \frac{n_1 Y}{n_{e1} n_{e2}} \right)^2 \right\},$$

where s_1 denotes the first-phase sample, $p(\phi)$ is the probability of drawing an empty sample, which is equal to

$$p(\phi) = p_1(\phi) + \sum_{s_1 \neq \phi} p_1(s_1) p_2(\phi),$$

and p_1 and p_2 denote respectively the sampling design for the first-phase and the second-phase sampling design conditional on the sample drawn in the first-phase.

Usually, population size is large and the first phase sample size is also large (compared to the second phase sample size). Thus we can safely assume $p_1(\phi) \doteq 0$ (compared to $p_2(\phi)$). For example, if we draw a first phase sample with expected sample size 50 out of a population of size 500, and then we draw a second phase sample with expected sample size 20 out of the first phase sample, all by using bionomial sampling, the inclusion probability in the first phase is 0.1 and the probability to draw an empty first phase sample is $(0.9)^{500}$; but the inclusion probability in the second phase is roughly .04 and the probability to draw an empty second phase sample is $(0.6)^{50}$. Notice that $(0.9)^{500} \doteq (0.3487)^{50} < (0.6)^{50}$. Thus, in most practical applications,

$$p_1(\phi) \doteq 0.$$

A variance estimator of \hat{Y}_2 can hence be easily given:

$$v_1(\hat{Y}_2) = p_2(\phi) \hat{Y}_2^2 + \frac{n_{e1} n_{e2}}{n_1 n_2} \sum_{i=1}^{n_2} (1 - \pi_{1i}) (y_i / \pi_{1i} - \hat{Y}_2 / n_{e1})^2 / \pi_{2i} + \frac{n_{e2}}{n_2} \left[\sum_{i=1}^{n_2} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{1i} \pi_{2i}} - \frac{n_1 \hat{Y}_2}{n_{e1} n_{e2}} \right)^2 \right]. \quad (5)$$

Estimator (5) should work well in usual applications. Sometimes when ocularly estimating net volume, however, the field worker may estimate that a tree has no value but turns out to be incorrect. Thus, some x_{2i} , hence π_{2i} , will be zero (in the simulations a small value is added to those so that $\pi_{2i} > 0$). In this case, a more stable term is needed to replace the last term in (5). Notice that

$$\frac{n_{e2}}{n_2} \sum_{i=1}^{n_2} \pi_{1i} (1 - \pi_{1i}) (y_i / \pi_{1i} - \hat{Y}_2 / n_{e1})^2 / \pi_{2i}$$

is an improved estimator of

$$\sum_{i=1}^{n_1} \pi_{1i} (1 - \pi_{1i}) (y_i / \pi_{1i} - \hat{Y}_2 / n_{e1})^2. \quad (6)$$

To ensure that the estimator does not become too large when one or more π_{2i} are close to zero, we use the following estimator

$$\left\{ \left[\sum_{i=1}^{n_2} \pi_{1i} (1 - \pi_{1i}) \left(\frac{y_i}{\pi_{1i}} - \frac{\hat{Y}_2}{n_{e1}} \right)^2 \right] / \left[\sum_{i=1}^{n_2} \pi_{2i} \right] \right\} n_{e2}. \quad (7)$$

If we consider x_{2i} as the auxiliary characteristic of $\pi_{1i} (1 - \pi_{1i}) (y_i/\pi_{1i} - \hat{Y}_2/n_{e1})^2$, then (7) is a ratio estimator of (6), since $\pi_{2i} \propto x_{2i}$ for $i = 1, \dots, n_1$. But since x_{2i} is not necessarily approximately proportional to $\pi_{1i} (1 - \pi_{1i}) (y_i/\pi_{1i} - \hat{Y}_2/n_{e1})^2$, (7) may not be a very efficient estimator of (6). The advantage of using (7) is that $\sum_{i=1}^{n_2} \pi_{2i}$ will not be close to zero, so that (7) will be stable.

This leads to the following variance estimator:

$$\begin{aligned} v_2(\hat{Y}_2) &= p_2(\phi) \hat{Y}_2^2 \\ &+ \frac{n_{e1}n_{e2}}{n_1} \left[\sum_{i=1}^{n_2} (1 - \pi_{1i}) \left(\frac{y_i}{\pi_{1i}} - \frac{\hat{Y}_2}{n_{e1}} \right)^2 \right] / \sum_{i=1}^{n_2} \pi_{2i} \\ &+ \frac{n_{e2}}{n_2} \sum_{i=1}^{n_2} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{1i}\pi_{2i}} - \frac{n_1\hat{Y}_2}{n_{e1}n_{e2}} \right)^2, \end{aligned} \quad (8)$$

which is less affected by small probabilities than (5) and hence is more stable. We will use (8) instead of (5) as a variance estimator of \hat{Y}_2 .

Let E_1 denote the expectation with respect to the first phase and E_2 denote the expectation with respect to the second phase. Since n_2 is the actual sample size and $E n_2 = E_1 E_2 n_2 = E_1 n_{e2}$, the adjusted estimator in PP sampling should be $E_1 n_{e2}/n_2 \hat{Y}_1$. But the quantity $E_1 n_{e2}$ is not available and is replaced by n_{e2} to obtain the following estimator:

$$\hat{Y}_3 = \frac{n_{e2}}{n_2} \hat{Y}_1. \quad (9)$$

\hat{Y}_3 should also have very small bias and the variance of \hat{Y}_3 is

$$\begin{aligned} V(\hat{Y}_3) &= p(\phi) Y^2 + \sum_{i=1}^N \frac{1 - \pi_{1i}}{\pi_{1i}} y_i^2 \\ &+ \sum_{s \neq \phi} p_1(s_1) \left\{ \sum_{i \in S_1} \pi_{2i} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{1i}\pi_{2i}} - \frac{Y}{n_{e2}} \right)^2 \right\}. \end{aligned}$$

A variance estimator of \hat{Y}_3 is

$$\begin{aligned} v(\hat{Y}_3) &= p_2(\phi) \hat{Y}_3^2 \\ &+ \frac{n_{e2}}{n_2} \left[\sum_{i=1}^{n_2} \pi_{2i} (1 - \pi_{1i}) \left(\frac{y_i}{\pi_{1i}\pi_{2i}} \right)^2 \right] \\ &+ \sum_{i=1}^{n_2} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{1i}\pi_{2i}} - \frac{\hat{Y}_3}{n_{e2}} \right)^2. \end{aligned} \quad (10)$$

Another possible estimator is based on the idea that we first want an efficient estimator of the first-phase information. This is accomplished by an analogous estimator to \hat{Y}_a in eq. (2):

$$\hat{Y}_a(2) = \sum_{i=1}^{n_2} (y_i/\pi_{2i}) n_{e2}/n_2 \quad \text{if } n_2 > 0.$$

This estimator can be expanded to estimate Y by dividing the first-phase sample by its probability of selection and we obtain

$$\hat{Y}_4 = \left[\hat{Y}_a(2) / \left\{ \prod_{i \in s} p_{1i} \prod_{j \notin s} (1 - p_{1j}) \right\} \right] / 2^{N-1}, \quad (11)$$

where $i \in s$ indicates that unit i is in the sample, $j \notin s$ indicates that j is not in the sample, $p_{1i} = n_{e1}x_{1i}/X_1$, and 2^{N-1} is the number of all samples.

The variance of \hat{Y}_4 is

$$\begin{aligned} V(\hat{Y}_4) &\doteq (2^{-2(N-1)}) \left[\sum_{s_1 \neq \phi} T(s_1)^2 / p_1(s_1) \right] - Y^2 \\ &+ (2^{-2(N-1)}) \sum_{s_1 \neq \phi} \left\{ \sum_{i \in S_1} \pi_{2i} (1 - \pi_{2i}) \right. \\ &\quad \left. \left[\frac{y_i}{\pi_{2i}} - \frac{1}{n_{e2}} T(s_1) \right]^2 + p_2(\phi) T(s_1)^2 \right\} / p_1(s_1), \end{aligned}$$

where $T(s_1)$ is the total of y over s_1 . It can be easily derived by using the formula

$$V(\hat{Y}_4) = V_1 E_2(\hat{Y}_4) + E_1 V_2(\hat{Y}_4),$$

and the variance given for \hat{Y}_a .

This estimator is expected to be highly unstable. A possible improvement is to condition the estimator on the actual sample size obtained, *i.e.*,

$$\hat{Y}_5 = \left[\frac{\hat{Y}_a(2)}{P_1(n_1)} \right] \cdot \left(\frac{N-1}{n_1-1} \right), \quad (12)$$

where $P_1(n_1)$ is the probability of drawing a first phase sample of size n_1 .

To compute this probability, let I_i be the random variable which is 1 if unit i is in the sample and 0 otherwise. Hence $n_1 = \sum_{i=1}^N I_i$, and

$$E(n_1) = n_{e1}, \text{Var}(n_1) = \sum_{i=1}^N \pi_{1i}(1 - \pi_{1i}) = d.$$

If

$$r = \frac{n_0 - n_e}{\sqrt{d}},$$

$$\phi(r) = (2\pi)^{-1/2} \exp\left[-\frac{1}{2}r^2\right],$$

$$f_m(r) = \left[\frac{1}{\sqrt{d}} \right] \phi(r) \left[1 + \sum_{j=1}^m p_j(r) \right],$$

where $P_j(r)$ are Edgeworth polynomials. Then

$$P_1(n_1) \doteq f_{m1}(r) \text{ and specifically, for } m = 2$$

$$\begin{aligned} P_1(n_1) \doteq f_2(r) &= \left[\frac{1}{\sqrt{d}} \right] \phi(r) \left[1 + \frac{1 - 2\bar{\pi}}{6\sqrt{d}} (r^3 - 3r) \right. \\ &\quad + \frac{1}{4!} \frac{1 - 6\pi(1 - \pi)}{d} (r^4 - 6r^2 + 3) \\ &\quad \left. + \frac{10}{6!} \frac{(1 - 2\bar{\pi})^2}{d} (r^6 - 15r^4 + 45r^2 - 15) \right], \end{aligned}$$

where

$$\bar{\pi} = \frac{\sum_{i=1}^N \pi_i^2(1 - \pi_i)}{\sum_{i=1}^N \pi_i(1 - \pi_i)}, \quad \overline{\pi(1 - \pi)} = \frac{\sum_{i=1}^N \pi_{1i}^2(1 - \pi_{1i})^2}{\sum_{i=1}^N \pi_{1i}(1 - \pi_{1i})} \quad (\text{Hájek 1981}).$$

\hat{Y}_4 and \hat{Y}_5 are only given for completeness. They are not considered further since both are unstable.

An alternative to \hat{Y}_4 and \hat{Y}_5 is to correct $\hat{Y}_a(2)$ using an expansion factor based on the information for covariate x_1 . These estimators are sensible if $\hat{Y}_a(2)/\sum_{i=1}^{n_1} x_{1i}$ is an approximately unbiased estimator of $R = Y/X_1$ which is true for binomial-Poisson (BP) but not for PP sampling. This fact is verified by simulation, but the reason why approximate unbiasedness holds for binomial-Poisson is that $\hat{Y}_a(2)/\sum_{i=1}^{n_1} x_{1i}$ under binomial sampling is similar to the ratio estimator under simple random sampling. Hence the following estimator is only appropriate for BP sampling.

$$\hat{Y}_6 = X_1 \left[\hat{Y}_a(2) / \sum_{i=1}^{n_1} x_{1i} \right]. \quad (13)$$

The variance of \hat{Y}_6 is

$$\begin{aligned} V(\hat{Y}_6) &= \frac{N^2}{n_{e1}^2} \sum_{i=1}^N (y_i - Rx_i)^2 \pi_{1i} \\ &\quad + E_1 \left\{ \frac{x_i}{n_1 \bar{x}_1 s_1} \left[\sum_{i=1}^{n_1} \pi_{2i} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{2i}} - \frac{n \bar{y}_{s1}}{n_{e2}} \right)^2 \right. \right. \\ &\quad \left. \left. + p_2(\phi) n_1^2 \bar{y}_{s1}^2 \right] \right\}. \end{aligned}$$

Another promising estimator is based on Srivastava's (1985) proposed unbiased estimator \hat{Y}_{sr1} based on the sample weight function concept. Srivastava and Ouyang (1992) developed a structure for the sample weight in order that \hat{Y}_{sr1} has zero variance at some points of the parameter space $\{y_1, \dots, y_N\}$. The sample weight function can use any information other than that given in a sample. Examples of this kind of information have been given in Srivastava and Ouyang (1992) and Ouyang and Schreuder (1992). If the information can be formulated as a model

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, N, \quad (14)$$

then the so called “generalized ratio estimator approximation” (Ouyang *et al.* (1992)) can be used which gives the following estimator of the population total:

$$\hat{Y}_7 = \left[\frac{\hat{Y}_1}{\sum_{i=1}^{n_2} y_i^* / (\pi_{1i} \pi_{2i})} \right] Y^*, \quad (15)$$

with $\hat{\alpha}$ and $\hat{\beta}$ weighted regression coefficients, and y_i^* calculated by $y_i^* = \hat{\alpha} + \hat{\beta}x_{1i}$ and $Y^* = \sum_{i=1}^N y_i^*$.

Note that \hat{Y}_7 is dependent on the model assumption.

6. SIMULATIONS

Simulation samples with first-and second-phase samples of expected sizes 50 and 20 in Poisson-Poisson and binomial-Poisson sampling were each drawn from three populations. Two populations were high-value fir, cedar and pine trees. Population 1, called BLM1 (Data from unpublished report “Comparison of volume estimates made by several timber measurement methods in western Oregon” by G. B. Hartman, Feb., 1971, Bureau of Land Management, Portland, Oregon), contained 331 trees and population 2, called BLM2, included 510 trees (Johnson and Hartman 1972). Measured variables on each tree were: net volume scaled (nvs), net volume dendrometered (nvd), and diameter at breast height (d). Here nvs ($= y$) is the variable of interest, $x_1 = d^2$ is used in the first phase of PP sampling and $x_2 = \text{nvd}$ is the more expensively but presumably additionally useful covariate obtained at the second level of PP sampling; 200,000 simulations were performed. Ideally, one would like the first- and second-level covariates to be relatively uncorrelated yet both highly correlated with y . These would be d^2 or nvd at the first phase and some measure of defect at the second-phase. Unfortunately, to do this in a satisfactory manner requires separating trees into a class where the field worker is comfortable estimating defect and another class for which he does not. This was not done for the available data. In BP sampling trees were selected with equal probabilities at the first phase and proportional to x_2 at the second phase. Population 3, a mapped data set, called Surinam, was also used since it was cleaner than the other populations in terms of having available more sensible variables for Poisson-Poisson sampling. The population consists of a 60-ha mapped Surinam forest for which only species and diameters were recorded (Schreuder *et al.* 1987). Tree heights and standing tree volumes for other species were superimposed on these trees as described in Schreuder *et al.* (1992). The resulting population consists of 5,525 trees for which tree diameter (d), height (h) and volume (v) were available. This yielded covariates $x_1 = h_h^2$ and

$x_2 =$ standing gross tree volume for PP sampling. For BP sampling x_2 was used at the second phase. Board foot volume (y) was also added to the data set. Included are 10 trees for which d^2h is large ($\geq 60,000$) but bd. ft. volume is essentially zero; 10,000 simulations were performed for the Surinam data. Results for BLM1, BLM2, and Surinam are given in Tables 1, 2 and 3 respectively.

Table 1

Simulation results for BLM1 ($N = 331$) population. 200,000 simulations were performed using $x_1 = D^2$ and $x_2 = \text{nvd}$ as covariates*

Estimator	Bias		SE		EASE	
	BP	PP	BP	PP	BP	PP
\hat{Y}_1	0.021	0.011	42.495	53.228		
\hat{Y}_2	-0.045	-0.770	37.272	48.219	97.787	97.806
\hat{Y}_3	-0.050	-0.777	39.819	49.349	97.492	96.763
\hat{Y}_6	0.012		39.992			
\hat{Y}_7	-0.036	3.650	18.881	21.885		

Table 2

Simulation results for BLM2 ($N = 510$) population. 200,000 simulations were performed using $x_1 = D^2$ and $x_2 = \text{nvd}$ as covariates*

Estimator	Bias		SE		EASE	
	BP	PP	BP	PP	BP	PP
\hat{Y}_1	0.146	0.059	95.708	62.500		
\hat{Y}_2	0.055	-0.424	90.247	55.876	100.325	98.583
\hat{Y}_3	0.050	-0.411	91.259	58.701	99.779	98.679
\hat{Y}_6	0.146		94.100			
\hat{Y}_7	0.486	4.391	26.788	19.855		

Table 3

Simulation results for Surinam ($N = 5,525$) population. 10,000 simulations were performed using $x_1 = D^2$ and $x_2 =$ ocular estimate of net volume*

Estimator	Bias		SE		EASE	
	BP	PP	BP	PP	BP	PP
\hat{Y}_1	0.764	0.364	25.709	25.924		
\hat{Y}_2	0.290	-0.402	15.636	10.845	97.492	97.37
\hat{Y}_3	0.019	-0.463	20.989	17.886	100.364	98.945
\hat{Y}_6	1.013		20.822			
\hat{Y}_7	2.277	2.426	22.428	17.397		

* All tables give bias and standard error (SE) expressed as a percentage of the population net volume. The estimated average standard error (EASE) is expressed as a percentage of the simulation standard error. Expected sample sizes are $n_{e1} = 50$ and $n_{e2} = 20$ for both binomial-Poisson (BP) and Poisson-Poisson (PP) sampling.

7. RESULTS AND DISCUSSION

For PP sampling \hat{Y}_2 is the most efficient estimator of the three (\hat{Y}_1 , \hat{Y}_2 , and \hat{Y}_3) relatively assumption-free estimators for BLM1 and BLM2; \hat{Y}_3 is slightly less efficient than \hat{Y}_2 . Note that \hat{Y}_7 is even more efficient than \hat{Y}_2 but \hat{Y}_7 has a serious bias in some cases. The variance estimators for \hat{Y}_2 and \hat{Y}_3 , $v(\hat{Y}_2)$ and $v(\hat{Y}_3)$, in eq. (8) and (10) are approximately unbiased.

For BP sampling, \hat{Y}_7 has negligible bias and the smallest standard error of all the estimators. \hat{Y}_2 is considerably less efficient than \hat{Y}_7 for BLM1 and BLM2 but more efficient than the other estimators. The variance estimators for both \hat{Y}_2 and \hat{Y}_3 are approximately unbiased.

Note for BLM1, BP sampling is always more efficient than PP sampling whereas for BLM2 PP sampling is more efficient with \hat{Y}_1 , \hat{Y}_2 and \hat{Y}_3 . This is because x_2 is not the logical variable to measure after the effect of x_1 is removed. Unfortunately a better variable to assess defect was not available for these data. For BLM1 x_2 did not but for BLM2 it did improve estimation.

For both PP and BP sampling, using population Surinam, \hat{Y}_2 is again the most efficient estimator of the three (\hat{Y}_1 , \hat{Y}_2 and \hat{Y}_3) relatively assumption-free estimators. \hat{Y}_3 is considerably less efficient than \hat{Y}_2 . \hat{Y}_7 is less efficient than \hat{Y}_2 and is substantially more biased for this population. $v(\hat{Y}_2)$ and $v(\hat{Y}_3)$ seems to be a approximately unbiased variance estimators for \hat{Y}_2 and \hat{Y}_3 . For this population PP sampling is more efficient than BP sampling with \hat{Y}_2 showing that in this case both $x_1 = d^2h$ and $x_2 =$ standing gross total volume are useful in sampling.

Actually, it is not surprising to see \hat{Y}_2 is the most efficient estimator, since it uses the most amount of information at both the design and estimation stages. Estimator \hat{Y}_7 tends to be even more efficient in terms of mean squared error, but with larger bias. This is because \hat{Y}_7 is based on the model given in equation (14). If the model is correct, \hat{Y}_7 should be preferred over \hat{Y}_2 , since \hat{Y}_7 incorporates even more information from the population. But otherwise, \hat{Y}_2 should be preferred. \hat{Y}_7 is not recommended if model (14) is not justified.

8. RECOMMENDATIONS

1. Both Poisson-Poisson and binomial-Poisson sampling are useful in practical forest sampling. With either procedure, estimator \hat{Y}_2 should be used. This estimator, with negligible bias and high efficiency, is analogous to the adjusted estimator \hat{Y}_a used in Poisson sampling and has a reliable variance estimator.
2. Estimator \hat{Y}_7 is considerably more efficient than \hat{Y}_2 for 2 populations but should not be used in preference to \hat{Y}_2 until it has been more fully investigated in additional studies. \hat{Y}_7 tends to be seriously biased in these simulations.

ACKNOWLEDGEMENT

We appreciate valuable comments by a referee.

REFERENCES

- BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.
- GOODMAN, L.A. (1949). On the estimation of the number of classes in a population. *Annals Mathematical Statistics*, 20, 572-579.
- GROSENBAUGH, L.R. (1964). Some suggestions for better sample-tree measurement. *Society of American Foresters. Proceedings*, 36-42.
- HAJAK, J. (1957). Some contributions to the theory of probability sampling. *Bulletin of the International Statistical Institute*, 36, 127-133.
- JOHNSON, F.A., and HARTMAN, G.B. (1972). Fall, buck and scale cruising. *Journal of Forestry*, 566-568.
- OUYANG, Z. (1990). Investigation of some estimators and strategies in sampling proposed by Srivastava. PhD thesis. Colorado State University Fort Collins, CO, 83.
- OUYANG, Z., and SCHREUDER, H.T. (1992). Srivastava estimation in forestry. Submitted to *Forest Science*.
- OUYANG, Z., SRIVASTAVA, J.N., and SCHREUDER, H.T. (1992). A general ratio estimator and its application in model based inference. *Annals. Institute of Statistical Mathematics*, (in press).
- SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- SCHREUDER, H.T., SEDRANSK, J., and WARE, K.D. (1968). Sampling and some alternatives, I. *Forest Science*, 14, 429-454.
- SCHREUDER, H.T., BANYARD, S.C., and BRINK, G.E. (1987). Comparison of three sampling methods in estimating stand parameters for a tropical forest. *Forest Ecology and Management*, 21, 119-128.
- SCHREUDER, H.T., and OUYANG, Z. (1992). Optimal sampling strategies for weighted linear regression estimation. *Canadian Journal of Forest Research*, 22, 239-247.
- SCHREUDER, H.T., OUYANG, Z., and WILLIAMS, M. (1992). Point-Poisson, point-pps, and modified point-pps sampling: Efficiency and variance estimation. *Canadian Journal of Forest Research*, (in press).
- SINGH, D., and SINGH, B.D. (1965). Some contributions to two-phase sampling. *Australian Journal of Statistics*, 7, 45-47.
- SRIVASTAVA, J.N. (1985). On a general theory of sampling, using experimental design. Concepts I: Estimation. *Bulletin of the International Statistical Institute*, 51, 1-16.
- SRIVASTAVA, J.N., and OUYANG, Z. (1992). Studies on a general estimator in sampling, utilizing extraneous information through a sampling weight function. *Journal of Statistical Planning and Inference*, 31, 177-196.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 8, Number 4, 1992

The Use of Composite Estimators with Two Stage Repeated Sample Design <i>D. Holt and T. Farver</i>	405
Nonresponse Adjustments for a Telephone Follow-up to a National In-Person Survey <i>Hüseyin Göksel, David R. Judkins and William D. Mosher</i>	417
Smoothing Variance Estimates for Price Indexes Over Time <i>Richard Valliant</i>	433
Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Operations <i>Michael F. Weeks</i>	445
 Miscellanea	
Training of African Statisticians at a Professional Level <i>James P.M. Ntozi</i>	467
Statistical Defensibility as Used by U.S. Department of Agriculture, National Agricultural Statistics Service <i>Rich Allen</i>	481
The United States Decennial Census: Problems, Possibilities and Prospects <i>William P. O'Hare</i>	499
Letters to the Editor	513
Special Notes	517
In Other Journals	519
Book Reviews	521
Editorial Collaborators	533
Index to Volume 8, 1992	539

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Division, Statistics Sweden, S-115 81 Stockholm, Sweden

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 42, No. 2, 1993

	<i>Page</i>
Stochastic ordering approach to off-line quality control <i>S. N. U. A. Kirmani and S. D. Peddada</i>	271
A three-state multiplicative model for rodent tumorigenicity experiments <i>J. C. Lindsey and L. M. Ryan</i>	283
Reallocation outliers in time series <i>L. S.-Y. Wu, J. R. M. Hosking and N. Ravishanker</i>	301
The shrinkage of point scoring methods <i>J. B. Copas</i>	315
Estimation of infant mortality rates categorized by social class for an Australian population <i>M. P. Quine and S. Quine</i>	333
Robust, smoothly heterogeneous variance regression <i>M. Cohen, S. R. Dalal and J. W. Tukey</i>	339
Modelling the relationship between crime count and observation period in prison inmates' self-report data <i>K. T. Hurrell</i>	355
Intervals which leave the minimum sum of absolute errors regression unchanged <i>S. C. Narula, V. A. Sposito and J. F. Wellington</i>	369
<i>General Interest Section</i>	
Interpretation of transformed axes in multivariate analysis <i>G. M. Arnold and A. J. Collins</i>	381
<i>Letters to the Editors</i>	401
<i>Book Reviews</i>	407
<i>Statistical Software Review</i>	
NANOSTAT	415
<i>Statistical Algorithms</i>	
AS 282 High breakdown regression and multivariate estimation <i>D. M. Hawkins and J. S. Simonoff</i>	423
AS 283 Rapid computation of the permutation paired and grouped <i>t</i> -tests <i>R. D. Baker and J. B. Tilbury</i>	432

Printed in Great Britain at the Alden Press, Oxford

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, I).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage pour la référence fois.)

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 42, No. 2, 1993

Page	
271	Stochastic ordering approach to off-line quality control <i>S. N. U. A. Kirmani and S. D. Peddada</i>
283	A three-state multiplicative model for rodent tumorigenicity experiments <i>J. C. Lindsey and L. M. Ryan</i>
301	Reallocation outliers in time series <i>L. S.-Y. Wu, J. R. M. Hosking and N. Ravishanker</i>
315	The shrinkage of point scoring methods <i>J. B. Copas</i>
333	Estimation of infant mortality rates categorized by social class for an Australian population <i>M. P. Quine and S. Quine</i>
339	Robust, smoothly heterogeneous variance regression <i>M. Cohen, S. R. Dalal and J. W. Tukey</i>
355	Modelling the relationship between crime count and observation period in prison inmates' self-report data <i>K. T. Hurrell</i>
369	Intervals which leave the minimum sum of absolute errors regression unchanged <i>S. C. Narula, V. A. Sposito and J. F. Wellington</i>
381	<i>General Interest Section</i> Interpretation of transformed axes in multivariate analysis <i>G. M. Arnold and A. J. Collins</i>
401	<i>Letters to the Editors</i>
407	<i>Book Reviews</i>
415	<i>Statistical Software Review</i> NANOSTAT
423	<i>Statistical Algorithms</i> AS 282 High breakdown regression and multivariate estimation <i>D. M. Hawkins and J. S. Simonoff</i>
432	AS 283 Rapid computation of the permutation paired and grouped <i>t</i> -tests <i>R. D. Baker and J. B. Tillybury</i>

Printed in Great Britain at the Alden Press, Oxford

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 8, Number 4, 1992

405	The Use of Composite Estimators with Two Stage Repeated Sample Design <i>D. Holt and T. Farver</i>
417	Nonresponse Adjustments for a Telephone Follow-up to a National In-Person Survey <i>Hüseyin Gökse, David R. Judkins and William D. Mosher</i>
433	Smoothing Variance Estimates for Price Indexes Over Time <i>Richard Valliant</i>
445	Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Operations <i>Michael F. Weeks</i>

Miscellanea

467	Training of African Statisticians at a Professional Level <i>James P.M. Niozi</i>
481	Statistical Defensibility as Used by U.S. Department of Agriculture, National Agricultural Statistics Service <i>Rich Allen</i>
499	The United States Decennial Census: Problems, Possibilities and Prospects <i>William P. O'Hare</i>

513	Letters to the Editor
517	Special Notes
519	In Other Journals
521	Book Reviews
533	Editorial Collaborators
539	Index to Volume 8, 1992

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Division, Statistics Sweden, S-115 81 Stockholm, Sweden

SÄRNDAAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.

SCHREUDER, H. T., SEDRANSK, J., et WARE, K. D. (1968). Sampling and some alternatives. *J. Forest Science*, 14, 429-454.

SCHREUDER, H. T., BANYARD, S. C., et BRINK, G. E. (1987). Comparison of three sampling methods in estimating stand parameters for a tropical forest. *Forest Ecology and Management*, 21, 119-128.

SCHREUDER, H. T., et OUYANG, Z. (1992). Optimal sampling strategies for weighted linear regression estimation. *Canadian Journal of Forest Research*, 22, 239-247.

SRIVASTAVA, J. N., et OUYANG, Z. (1992). Studies on a general estimator in sampling, utilizing extraneous information through a sampling weight function. *Journal of Statistical Planning and Inference*, 31, 177-196.

SINGH, D., et SINGH, B. D. (1965). Some contributions to two-phase sampling. *Australian Journal of Statistics*, 7, 45-47.

SRIVASTAVA, J. N. (1985). On a general theory of sampling, using experimental design. Concepts I: Estimation. *Bulletin de l'Institut International de Statistique*, 51, 1-16.

SCHREUDER, H. T., OUYANG, Z., et WILLIAMS, M. (1992). Point-Poisson, point-pps, and modified point-pps sampling: Efficiency and variance estimation. *Canadian Journal of Forest Research*, (à paraître).

puisque X_1 incorpore encore plus d'information relative à la population. Sinon, X_2 constitue le meilleur choix. X_1 n'est pas recommandé si l'emploi du modèle (14) n'est pas justifié.

8. RECOMMANDATIONS

1. L'échantillonnage Poisson-Poisson et l'échantillonnage binomial-Poisson conviennent tous deux à l'exécution pratique d'échantillonnages en forêt. Dans les deux cas, l'estimateur \hat{X}_2 devrait être utilisé. Cet estimateur, qui présente un biais négligeable et une efficacité élevée, est analogue à l'estimateur modifié \hat{X}_n utilisé dans l'échantillonnage de Poisson et possède un estimateur de variance fiable.
2. L'estimateur \hat{X}_1 est beaucoup plus efficace que \hat{X}_2 pour deux populations, mais ne devrait pas être préféré à \hat{X}_2 avant d'avoir été plus abondamment analysé dans le cadre d'autres études. \hat{X}_1 tend à afficher un biais élevé dans les simulations effectuées.

REMERCIEMENTS

Nous apprécions les commentaires valables qui ont été formulés par un arbitre.

BIBLIOGRAPHIE

- BREWER, K.R.W., et HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.
- GOODMAN, L.A. (1949). On the estimation of the number of classes in a population. *Annals Mathematical Statistics*, 20, 572-579.
- GROSENBAGH, L.R. (1964). Some suggestions for better sample-tree measurement. *Society of American Foresters. Proceedings*, 36-42.
- HAIK, J. (1957). Some contributions to the theory of probability sampling. *Bulletin de l'Institut International de Statistique*, 36, 127-133.
- JOHNSON, F.A., et HARTMAN, G.B. (1972). Fall, buck and scale cruising. *Journal of Forestry*, 566-568.
- OUYANG, Z. (1990). Investigation of some estimators and strategies in sampling proposed by Srivastava. Thèse de doctorat, Colorado State University Fort Collins, CO, 83.
- OUYANG, Z., et SCHREUDER, H.T. (1992). Srivastava estimation in forestry. Soumis à *Forest Science*.
- OUYANG, Z., SRIVASTAVA, J.N., et SCHREUDER, H.T. (1992). A general ratio estimator and its application in model based inference. *Annals. Institute of Statistical Mathematics*, (à paraître).

covariables $x_1 = h_1^2$ et $x_2 =$ volume brut des arbres sur pied pour l'échantillonnage PP. Dans le cas de l'échantillonnage BP, x_2 a été utilisé à la deuxième phase. Le volume en pieds-planche (v) a été ajouté à l'ensemble de données. La population comprend 10 arbres pour lesquels d^2h est élevé ($\geq 60,000$), mais pour lesquels le volume en pieds-planche est essentiellement nul; les données sur la population Surinam ont été soumises à 10,000 simulations. Les résultats relatifs aux populations BLM1, BLM2 et Surinam sont donnés aux tableaux 1, 2 et 3 respectivement.

7. RÉSULTATS ET ANALYSE

En ce qui concerne l'échantillonnage PP, \hat{X}_2 est l'estimateur le plus efficace parmi les trois estimateurs (\hat{X}_1 , \hat{X}_2 , et \hat{X}_3) qui, relativement, ne dépendent pas d'hypothèses, pour BLM1 et BLM2; \hat{X}_3 est légèrement moins efficace que \hat{X}_2 . Notons que \hat{X}_1 est encore plus efficace dans certains cas, mais \hat{X}_1 souffre d'un biais prononcé dans certains cas. Les estimateurs de variance pour \hat{X}_2 et \hat{X}_3 , $v(\hat{X}_2)$ et $v(\hat{X}_3)$, donnés aux équations (8) et (10) sont approximativement sans biais.

Pour l'échantillonnage BP, \hat{X}_1 affiche un biais négligable et la plus faible erreur-type de tous les estimateurs. \hat{X}_2 est beaucoup moins efficace que \hat{X}_1 pour BLM1 et BLM2, mais plus efficace que les autres estimateurs. Les estimateurs de variance tant pour \hat{X}_2 que pour \hat{X}_3 sont approximativement sans biais.

Notons que pour BLM1, l'échantillonnage BP est toujours plus efficace que l'échantillonnage PP, tandis que pour BLM2, l'échantillonnage PP est plus efficace avec \hat{X}_1 , \hat{X}_2 et \hat{X}_3 . Cela tient au fait que x_2 n'est pas la variable logique à mesurer une fois que l'effet de x_1 est retranché. Malheureusement, nous ne disposons pas, pour ces données, d'une meilleure variable d'évaluation du volume inexploitable. Pour BLM1, x_2 n'a pas amélioré l'estimation, mais pour BLM2, il a produit une amélioration.

Tant pour l'échantillonnage PP que pour l'échantillonnage BP portant sur la population Surinam, \hat{X}_2 est encore l'estimateur le plus efficace parmi les trois estimateurs (\hat{X}_1 , \hat{X}_2 et \hat{X}_3) qui, relativement, ne dépendent pas d'hypothèses. \hat{X}_3 est beaucoup moins efficace que \hat{X}_2 . \hat{X}_1 est moins efficace que \hat{X}_2 et est beaucoup plus biaisé pour cette population. $v(\hat{X}_2)$ et $v(\hat{X}_3)$ semblent être des estimateurs de variance approximativement sans biais pour \hat{X}_2 et \hat{X}_3 . Pour cette population, l'échantillonnage PP est plus efficace que l'échantillonnage BP, et \hat{X}_2 montre que dans ce cas, tant $x_1 = d^2h$ que $x_2 =$ volume brut total des arbres sur pied sont utiles à l'échantillonnage.

En fait, il n'est pas surprenant de constater que \hat{X}_2 est l'estimateur le plus efficace, car c'est celui qui utilise le plus d'information, tant au niveau du plan d'échantillonnage qu'à celui de l'estimation. L'estimateur \hat{X}_1 tend à être encore plus efficace du point de vue de l'erreur quadratique moyenne, mais avec un biais plus élevé. Cela tient au fait que \hat{X}_1 se fonde sur le modèle donné à l'équation (14). Si le modèle est juste, \hat{X}_1 devrait être préféré à \hat{X}_2 .

seulement les espèces et les diamètres ont été relevés (Schreuder et coll. 1987). Des hauteurs d'arbres et des volumes d'arbres sur pied pour d'autres espèces ont été attribués à ces arbres, de la façon décrite dans Schreuder et coll. (1992). La population résultante comprend 5,525 arbres pour lesquels on dispose du diamètre (d), de la hauteur (h) et du volume (v). Ces valeurs donnent les

Table 1

Résultats des simulations pour la population BLM1 ($N = 331$). 200,000 simulations ont été effectuées avec $x_1 = D_2^*$ et $x_2 = \text{nvd}$ comme covariables*

Estimateur	Biases			Erreur-type			ETME
	BP	PP	BP	BP	PP	BP	PP
Y_1	0.021	0.011	42.495	53.228			
Y_2	-0.045	-0.770	37.272	48.219	97.787	97.806	
Y_3	-0.050	-0.777	39.819	49.349	97.492	96.763	
Y_6	0.012		39.992				
Y_7	-0.036	3.650	18.881	21.885			

Table 2

Résultats des simulations pour la population BLM2 ($N = 510$). 200,000 simulations ont été effectuées avec $x_1 = D_2^*$ et $x_2 = \text{nvd}$ comme variables*

Estimateur	Biases			Erreur-type			ETME
	BP	PP	BP	BP	PP	BP	PP
Y_1	0.146	0.059	95.708	62.500			
Y_2	0.055	-0.424	90.247	55.876	100.325	98.583	
Y_3	0.050	-0.411	91.259	58.701	99.779	98.679	
Y_6	0.146		94.100				
Y_7	0.486	4.391	26.788	19.855			

Table 3

Résultats des simulations pour la population Surinam ($N = 5,525$). 10,000 simulations ont été effectuées $x_1 = D_2^*$ et $x_2 = \text{estimation visuelle du volume net}$ *

Estimateur	Biases			Erreur-type			ETME
	BP	PP	BP	BP	PP	BP	PP
Y_1	0.764	0.364	25.709	25.924			
Y_2	0.290	-0.402	15.636	10.845	97.492	97.37	
Y_3	0.019	-0.463	20.989	17.886	100.364	98.945	
Y_6	1.013		20.822				
Y_7	2.277	2.426	22.428	17.397			

* Tous les tableaux donnent le biais et l'erreur-type sous forme du pourcentage du volume net pour la population. L'erreur-type moyenne estimative (ETME) est exprimée en pourcentage de l'erreur-type de simulation. Les espérances des tailles d'échantillon sont $n_{g1} = 50$ et $n_{g2} = 20$ tant pour l'échantillonnage binomial-Poisson (BP) que pour l'échantillonnage Poisson-Poisson (PP).

avec les coefficients de régression pondérés \hat{a} et $\hat{\beta}$, et y_i^* donné par $y_i^* = \hat{a} + \hat{\beta}x_{i1}$ et $Y^* = \sum_{i=1}^N y_i^*$.
Notons que Y_7 est tributaire de l'hypothèse du modèle.

6. SIMULATIONS

Des échantillons de simulation de première et de deuxième matique 50 et 20 unités, constitués selon l'échantillonnage Poisson-Poisson et l'échantillonnage binomial-Poisson, ont été prélevés pour trois populations. Deux populations étaient formées de sapins, de thuyas et de pins à valeur élevée. La population 1, appelée BLM1 (données du rapport non publié "Comparaison of volume estimates made by several timber measurement methods in western Oregon", par G.B. Hartman, février 1971. Bureau of Land Management, Portland, Oregon), contenait 331 arbres et la population 2, appelée BLM2, comprenait 510 arbres (Johnson et Hartman 1972). Les variables mesurées sur chaque arbre étaient les suivantes: volume net cubé (nvs), volume net mesuré au dendromètre (nvd) et diamètre à hauteur de poitrine (d). Ici, nvs ($= y$) est la variable à l'étude, $x_1 = d^2$ est utilisé à la première phase de l'échantillonnage PP et $x_2 = \text{nvd}$ est la covariable plus coûteuse à évaluer, mais vraisemblablement plus utile, obtenue à la deuxième phase de l'échantillonnage PP; 200,000 simulations ont été effectuées. Idéalement, il serait souhaitable que les covariables de la première et de la deuxième phase soient relativement non corrélées entre elles, tout en étant l'une et l'autre en étroite corrélation avec y . Ces variables seraient d^2 ou nvd à la première phase et une mesure du volume inexploitable à la deuxième phase. Malheureusement, il faudrait que les arbres soient séparés en une classe pour laquelle l'observateur sur place est en mesure d'estimer efficacement le volume inexploitable, et en une autre classe pour laquelle il ne l'est pas. Cela n'a pas été fait dans le cas des données disponibles. Dans l'échantillonnage BP, les arbres ont été sélectionnés avec probabilité égale à la première phase et avec probabilité proportionnelle à x_2 à la deuxième phase. La population 3, un ensemble de données cartographiques appelé Surinam, a également été utilisée car elle offrait une meilleure qualité que les autres populations sur le plan de la sensibilité des variables se prêtant à l'échantillonnage Poisson-Poisson. La population est constituée d'une forêt cartographiée de 60 hectares du Surinam pour laquelle

où $T(s_1)$ est le total de y sur s_1 . Elle peut être facilement obtenue au moyen de la formule

$$V(\hat{Y}_q) = V_1 E_2(\hat{Y}_q) + E_1 V_2(\hat{Y}_q),$$

et de la variance données pour \hat{Y}_q .

On peut s'attendre à ce que cet estimateur soit très

instable. Une amélioration possible consiste à assujettir l'estimateur à une condition, soit la taille réelle de l'échantillon obtenu, c'est-à-dire

$$\hat{Y}_5 = \left[\hat{Y}_q(2) \right] /$$

$$\left\{ \left[\prod_{i=1}^{n_1} \pi_{1i} \prod_{j \neq s} (1 - \pi_{1j}) \right] \frac{P_1(n_1)}{\binom{N-1}{n_1-1}} \right\}. \quad (12)$$

où $P_1(n_1)$ est la probabilité de tirer un échantillon de première phase de taille n_1 .

Pour calculer cette probabilité, définissons I_i comme la variable aléatoire égale à 1 si l'unité i appartient à l'échantillon, et égale à 0 sinon. Ainsi, $n_1 = \sum_{i=1}^N I_i$, et

$$E(n_1) = n_{e1}, \text{Var}(n_1) = \sum_{i=1}^N \pi_{1i}(1 - \pi_{1i}) = d.$$

Si

$$r = \frac{n_0 - n_e}{n_e},$$

$$\phi(r) = (2\pi)^{-1/2} \exp\left[-\frac{1}{2}r^2\right],$$

$$f_m(r) = \left[\frac{1}{\sqrt{d}} \phi(r) \right] \left[1 + \sum_{m=1}^j P_j(r) \right],$$

où les $P_j(r)$ sont des polynômes de Edgeworth. On a alors

$$P_1(n_1) = f_{m1}(r) \text{ et, plus précisément, pour } m = 2$$

$$P_1(n_1 = f_2(r)) = \left[\frac{1}{\sqrt{d}} \phi(r) \right] \left[1 + \frac{1}{1 - 2\pi} \frac{6\sqrt{d}}{(r^3 - 3r)} \right]$$

$$+ \frac{1}{4} \frac{1}{1 - 6\pi(1 - \pi)} (r^4 - 6r^2 + 3)$$

$$+ \frac{6i}{10} \frac{d}{(1 - 2\pi)^2} (r^6 - 15r^4 + 45r^2 - 15) \left. \right].$$

où

$$\pi = \frac{\sum_{i=1}^N \pi_2^i (1 - \pi_i)}{\sum_{i=1}^N \pi_2^i (1 - \pi_{1i})^2} = \frac{\sum_{i=1}^N \pi_i (1 - \pi)}{\sum_{i=1}^N \pi_{1i} (1 - \pi_{1i})}.$$

(Hájek 1981).

\hat{Y}_4 et \hat{Y}_5 sont donnés seulement par souci d'exhaustivité. Nous ne nous y intéresserons plus, car ils sont tous deux instables.

Une solution de rechange pour \hat{Y}_4 et \hat{Y}_5 consiste à corriger $\hat{Y}_q(2)$ au moyen d'un facteur d'expansion basé sur l'information relative à la covariable x_1 . Ces estimateurs sont raisonnables si $\hat{Y}_q(2) / \sum_{i=1}^{n_1} x_{1i}$ est un estimateur approximativement sans biais de $R = Y/X_1$, ce qui est vrai pour l'échantillonnage binomial-Poisson (BP), mais pas pour l'échantillonnage PP. Ce fait est vérifié par simulation, mais la raison pour laquelle l'estimateur est approximativement sans biais dans le cas de l'échantillonnage binomial-Poisson est que $\hat{Y}_q(2) / \sum_{i=1}^{n_1} x_{1i}$, pour un échantillonnage binomial, est semblable à l'estimateur par quotient utilisé dans l'échantillonnage aléatoire simple. Par conséquent, l'estimateur suivant ne convient qu'à l'échantillonnage BP.

$$\hat{Y}_6 = X_1 \left[\hat{Y}_q(2) / \sum_{i=1}^l x_{1i} \right]. \quad (13)$$

La variance de \hat{Y}_6 est

$$V(\hat{Y}_6) = \frac{N^2}{n_{e1}^2} \sum_{i=1}^N (Y_i - R X_i)^2 \pi_{1i}$$

$$+ E_1 \left\{ \frac{X_i}{n_1} \sum_{i=1}^l \pi_{2i} (1 - \pi_{2i}) \left(\frac{Y_i}{Y_l} - \frac{n_{e2}}{n_{e1}^2} \right)^2 \right\}.$$

Un autre estimateur prometteur se fonde sur l'estimateur sans biais \hat{Y}_{sr1} proposé par Srivastava (1985), d'après le concept de fonction de poids d'échantillon. Srivastava et Ouyang (1992) ont élaboré une structure pour le poids d'échantillon de telle sorte que \hat{Y}_{sr1} a une variance nulle à certains points de l'espace des paramètres $\{Y_1, \dots, Y_N\}$. La fonction de poids d'échantillon peut utiliser toute information autre que celle fournie dans un échantillon. Des exemples de ce genre d'information ont été donnés dans Srivastava et Ouyang (1992) et dans Ouyang et Schreuder (1992). Si l'information peut être présentée sous forme d'un modèle

$$Y_i = \alpha + \beta x_i + e_i, i = 1, \dots, N, \quad (14)$$

seront égaux à zéro (dans les simulations, une faible valeur est ajoutée à ces mesures de sorte que $\pi_{2i} > 0$). Dans ce cas, il faut un terme plus stable pour remplacer le dernier terme dans (5). Notons que

$$n_{e2} \sum_{n_2} \pi_{1i} (1 - \pi_{1i}) (y_i / \pi_{1i} - x_{2i} / n_{e1})^2 / \pi_{2i}$$

est un estimateur amélioré de

$$\sum_{n_1} \pi_{1i} (1 - \pi_{1i}) (y_i / \pi_{1i} - x_{2i} / n_{e1})^2. \quad (6)$$

Pour veiller à ce que l'estimateur ne devienne pas trop élevé quand un ou plusieurs π_{2i} sont voisins de zéro, nous utilisons l'estimateur suivant

$$\left\{ \left[\sum_{n_2} \pi_{1i} (1 - \pi_{1i}) \left(\frac{y_i}{\pi_{1i}} - \frac{x_{2i}}{n_{e1}} \right)^2 \right] / \right.$$

$$\left. \left[\sum_{n_2} \pi_{2i} \right] \right\} n_{e2}. \quad (7)$$

Si nous considérons x_{2i} comme la caractéristique auxiliaire de $\pi_{1i} (1 - \pi_{1i}) (y_i / \pi_{1i} - x_{2i} / n_{e1})^2$, l'expression (7) est un estimateur par quotient de (6), puisque π_{2i} a x_{2i} pour $i = 1, \dots, n_1$. Mais puisque x_{2i} n'est pas nécessairement approximativement proportionnel à $\pi_{1i} (1 - \pi_{1i}) (y_i / \pi_{1i} - x_{2i} / n_{e1})^2$, il se peut que (7) ne soit pas un estimateur très efficace de (6). L'avantage d'utiliser (7) est que $\sum_{n_2} \pi_{2i}$ ne sera pas voisin de zéro, de sorte que (7) sera stable.

Il en découle l'estimateur de variance suivant:

$$v_2(\hat{Y}_2) = p_2(\phi) \hat{Y}_2^2$$

$$+ \frac{n_{e1} n_{e2}}{n_1} \left[\sum_{n_2} \pi_{1i} (1 - \pi_{1i}) \left(y_i / \pi_{1i} - \frac{x_{2i}}{n_{e1}} \right)^2 \right] / \sum_{n_2} \pi_{2i} + \frac{n_{e2}}{n_2} \sum_{n_2} \pi_{2i} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{2i}} - \frac{n_1 \hat{Y}_2^2}{n_{e1} n_{e2}} \right)^2, \quad (8)$$

qui est moins intensément touché par les probabilités faibles que (5) et qui est donc plus stable. Nous utiliserons (8) plutôt que (5) comme estimateur de la variance de \hat{Y}_2 .

Supposons que E_1 désigne l'espérance s'appliquant à la première phase et E_2 l'espérance s'appliquant à la deuxième phase. Puisque n_2 est la taille réelle de l'échantillon et que $E n_2 = E_1 E_2 n_2 = E_1 n_{e2}$, l'estimateur modifié dans l'échantillonnage PP devrait être $E_1 n_{e2} / n_2 \hat{Y}_1$. Mais la quantité $E_1 n_{e2}$ n'est pas disponible et est remplacée par n_{e2} , ce qui donne l'estimateur suivant:

$$V(\hat{Y}_3) = (2^{-2(N-1)}) \left[\sum_{s_1} T(s_1)^2 / p_1(s_1) \right] - Y_2^2 + (2^{-2(N-1)}) \sum_{s_1 \neq \phi} \pi_{2i} (1 - \pi_{2i}) \left[\frac{y_i}{\pi_{2i}} - \frac{1}{n_{e1}} T(s_1)^2 + p_2(\phi) T(s_1)^2 / p_1(s_1) \right]$$

La variance de \hat{Y}_4 est

$$\hat{Y}_4 = \left[\hat{Y}_a(2) \right] / \left\{ \prod_{n_1} \prod_{n_2} p_{1i} \prod_{j \in s_1} (1 - p_{1j}) \right\} / 2^{N-1}, \quad (11)$$

Cet estimateur peut être étendu à l'estimation de Y_j à cette fin, nous divisons l'échantillon de première phase par sa probabilité de sélection, et nous obtenons

$$\hat{Y}_a(2) = \sum_{n_2} (y_i / \pi_{2i}) n_{e2} / n_2 \quad \text{si } n_2 > 0.$$

Un autre estimateur possible découle de l'idée que nous voulons d'abord un estimateur efficace de l'information de la première phase. À cette fin, nous utilisons un estimateur analogue à \hat{Y}_a donné à l'équation (2):

$$+ \sum_{n_2} \pi_{2i} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{2i}} - \frac{n_{e2}}{n_2} \right)^2. \quad (10)$$

$$v(\hat{Y}_3) = p_2(\phi) \hat{Y}_3^2$$

Un estimateur de la variance de \hat{Y}_3 est donné par

$$V(\hat{Y}_3) = p(\phi) Y_2^2 + \sum_{n_1} \frac{\pi_{1i}}{1 - \pi_{1i}} y_i^2 + \sum_{s_1} \pi_{2i} (1 - \pi_{2i}) \left(\frac{y_i}{\pi_{2i}} - \frac{n_{e2}}{n_2} \right)^2.$$

de \hat{Y}_3 est

$$\hat{Y}_3 = \frac{n_{e2}}{n_2} \hat{Y}_1. \quad (9)$$

\hat{Y}_3 devrait aussi être très faiblement biaisé et la variance

5. THÉORIE

Pour l'échantillonnage de Poisson, l'estimateur

$$\bar{Y}_n = \sum_{i=1}^n Y_i / \pi_i(P), \quad (1)$$

est non biaisé mais très inefficace et devrait être remplacé par l'estimateur suivant, approximativement sans biais (Grosenbaugh 1964):

$$\bar{Y}_a = \begin{cases} n_e \bar{Y}_n & \text{si } n > 0 \\ 0 & \text{si } n = 0 \end{cases} \quad (2)$$

$$= 0 \quad \text{si } n = 0.$$

La variance de \bar{Y}_a , donnée par Brewer et Hanif (1983),

est

$$V(\bar{Y}_a)$$

$$= \sum_{i=1}^n \pi_i(P) [1 - \pi_i(P)] \left[\frac{\pi_i(P)}{Y_i} - \frac{n_e}{Y} \right]^2 + p_0 Y^2,$$

où $p_0 = P(n = 0)$.

Pour l'échantillonnage Poisson-Poisson (PP), un estimateur de Y analogue à \bar{Y}_n ci-dessus est l'estimateur

sans biais

$$\bar{Y}_1 = \sum_{i=1}^{n_2} Y_i / \pi_i. \quad (3)$$

Cet estimateur peut être terriblement inefficace, comme il en a été fait mention pour \bar{Y}_n dans le cas de l'échantillonnage de Poisson (Schreuder et coll. 1968).

La variance de \bar{Y}_1 peut être exprimée au moyen des formules générales élaborées par Särndal et Swensson (1987) pour l'estimation non biaisée dans les plans à échantillon-

nage double:

$$V(\bar{Y}_1)$$

$$= \sum_{i=1}^N \left(\frac{1 - \pi_{1i}}{\pi_{1i}} \right) Y_i^2 + E_1 \left\{ \sum_{i=1}^n \left(\frac{1 - \pi_{2i}}{1 - \pi_{2i}} \right) \left(\frac{\pi_{1i}}{Y_i} \right)^2 \right\},$$

où E_1 désigne l'espérance s'appliquant à l'échantillon de première phase. Puisque \bar{Y}_1 n'est pas efficace, nous ne donnons pas l'estimateur de sa variance. Comme pour l'estimateur modifié plus efficace présenté dans le cas de l'échantillonnage de Poisson, nous avons l'estimateur approximativement non biaisé suivant

$$\bar{Y}_2 = \sum_{i=1}^{n_2} Y_i / \pi_i^* = \bar{Y}_1 (n_{e1}/n_1) (n_{e2}/n_2). \quad (4)$$

La variance de \bar{Y}_2 est:

$$V(\bar{Y}_2) = p(\phi) Y_2 + \sum_{i=1}^N \pi_{1i} (1 - \pi_{1i}) \left(\frac{\pi_{1i}}{Y_i} - \frac{\pi_{e1}}{Y} \right)^2 + \sum_{s_1 \neq \phi} p_1(s_1) \left\{ \sum_{i=1}^{n_2} \pi_{2i} (1 - \pi_{2i}) \left(\frac{\pi_{1i} \pi_{2i}}{Y_i} - \frac{n_{e1} n_{e2}}{n_1 Y} \right)^2 \right\},$$

où s_1 dénote l'échantillon de première phase, $p(\phi)$ est la probabilité de tirer un échantillon vide, qui est égale à

$$p(\phi) = p_1(\phi) + \sum_{s_1 \neq \phi} p_1(s_1) p_2(\phi),$$

et p_1 et p_2 dénotent respectivement le plan d'échantillonnage de la première phase, et le plan d'échantillonnage de la deuxième phase étant donné l'échantillon tiré à la première phase.

Habituellement, la taille de la population est élevée et l'échantillon de première phase est aussi de grande taille (comparativement à l'échantillon de deuxième phase). Par conséquent, nous pouvons supposer sans risque que $p_1(\phi) \approx 0$ (comparativement à $p_2(\phi)$). Par exemple, si nous prélevons un échantillon de première phase pour lequel l'espérance de la taille est de 50 unités sur une population de taille 500, et que nous tirons ensuite, parmi l'échantillon de première phase, un échantillon de deuxième phase pour lequel l'espérance de la taille est de 20 unités, en utilisant dans tous les cas l'échantillonnage binomial, la probabilité d'inclusion à la première phase est de 0,1 et la probabilité de tirer un échantillon de première phase vide est de $(0,9)^{500}$; toutefois, la probabilité d'inclusion à la deuxième phase est d'environ 0,04 et la probabilité de prélever un échantillon de deuxième phase vide est de $(0,6)^{50}$. Notons que $(0,9)^{500} \approx (0,3487)^{50} < (0,6)^{50}$. Ainsi, dans la plupart des applications pratiques,

$$p_1(\phi) \approx 0.$$

Un estimateur de la variance de \bar{Y}_2 peut donc être facilement formulé:

$$v_1(\bar{Y}_2) = p_2(\phi) Y_2^2$$

$$+ \frac{n_{e1} n_{e2}}{n_1 n_2} \sum_{i=1}^N (1 - \pi_{1i}) (Y_i / (\pi_{1i} - \bar{Y}_2 / n_{e1}))^2 / \pi_{2i}$$

$$+ \frac{n_{e2}}{n_2} \left[\sum_{i=1}^n (1 - \pi_{2i}) \left(\frac{\pi_{1i} \pi_{2i}}{Y_i} - \frac{n_{e1} \bar{Y}_2}{n_1 n_2} \right)^2 \right]. \quad (5)$$

L'estimateur (5) devrait bien fonctionner dans les situations courantes. Toutefois, il peut arriver que l'observateur sur place, lorsqu'il estime visuellement le volume net, juge qu'un arbre est sans valeur mais commette ainsi une erreur. Par conséquent, certains x_{2i} , et donc certains π_{2i} ,

3. MÉTHODES D'ÉCHANTILLONNAGE

La région 6 du United States Forest Service (Wendall L. Jones – communication personnelle) utilise la méthode d'échantillonnage des chargements de camions suivante: au moment où les camions s'arrêtent à l'usine, une technique d'échantillonnage binomiale est utilisée pour sélectionner au hasard les camions à inclure dans l'échantillon, avec $p = 0,10$ par exemple. Le volume des chargements sélectionnés est mesuré. Une difficulté de cette méthode est qu'il y a parfois de longues suites de chargements non prélevés dans l'échantillon. Comme il a été signalé à l'un des auteurs, cette situation était jugée hautement indésirable d'un point de vue pratique. Une autre méthode, qui devrait diminuer la fréquence des longues suites de chargements non sélectionnés, et qui pourrait être plus efficace, consiste à utiliser plutôt l'échantillonnage binomial-Poisson, de la façon suivante: Un échantillonnage binomial est effectué avec un p plus élevé (par exemple $p = 0,30$), et le volume des chargements ainsi sélectionnés est estimé visuellement par le mesureur. Puis, un sous-échantillon de Poisson de ces chargements est prélevé avec probabilité proportionnelle aux volumes estimés, et chacun des chargements sélectionnés à cette phase fait l'objet d'une mesure du volume. C'est ce que l'on appelle l'échantillonnage binomial-Poisson.

Pour les peuplements forestiers à valeur élevée de la région du Pacifique nord-ouest des États-Unis, des estimations très précises du volume net, c'est-à-dire du volume exploitable, sont souvent nécessaires. L'abattage réel et la mesure destructive des arbres de l'échantillon constituent la méthode la plus fiable de détermination du volume net, soit le volume total moins le volume inexploitable (Johnson et Hartman 1972). L'échantillonnage Poisson-Poisson peut représenter un plan d'échantillonnage approprié dans cette situation. En voici les étapes:

1. Sélectionner, par un échantillonnage de Poisson, n_1 arbres parmi les N arbres formant la population, avec probabilité de sélection proportionnelle à une estimation du volume brut, par exemple $x_1 = \text{diamètre à hauteur de poitrine élevé au carré } (d^2)$. Avec l'échantillonnage de Poisson, la taille réelle de l'échantillon est aléatoire, disons n_1 avec $E(n_1) = n_{e1}$. Estimer ensuite visuellement, par exemple, $x_2 = \text{volume net visuel}$.
2. Sélectionner n_2 arbres parmi l'échantillon de n_1 arbres, avec probabilité proportionnelle à x_2 , par un échantillonnage de Poisson. Ici, $E(n_2) = n_{e2}$ est l'espérance de la taille de l'échantillon à la deuxième phase.

Les n_2 arbres sélectionnés sont ensuite abattus et détruits pour permettre la mesure des volumes brut, net et inexploitable. Pour assurer le maximum d'efficacité de l'inventaire et de l'exécution du travail, il est sans doute préférable de réaliser les deux phases au même moment et de marquer les n_2 arbres de l'échantillon au moment de l'inventaire. L'évaluation du volume exploitable pour ces n_2 arbres est effectuée plus tard par une équipe différente, ou encore les arbres de l'échantillon sont transportés dans une usine où ils seront réellement transformés en produits du bois. L'échantillonnage binomial-Poisson est un cas spécial de cette dernière méthode. (Si la deuxième phase est réalisée

4. NOTATION

à un autre moment que la première, une liste d'unités d'échantillonnages est disponible pour l'exécution de cette deuxième phase et une méthode avec ppt et taille d'échantillon fixe devrait être utilisée au lieu de l'échantillonnage de Poisson. Cette façon de procéder est généralement inefficace, car il faut alors se rendre deux fois sur le terrain.)

N = taille de la population (inconnue jusqu'à ce que l'échantillonnage soit terminé).

n_e = espérance de la taille de l'échantillon dans l'échantillonnage de Poisson à une seule phase.

n = taille réelle de l'échantillon dans l'échantillonnage de Poisson à une seule phase.

n_{e1} = espérance de la taille de l'échantillon à la première phase de l'échantillonnage de Poisson à deux phases.

n_2 = taille réelle de l'échantillon de deuxième phase de l'échantillonnage de Poisson à deux phases.

Y = volume exploitable total de la population (valeur à estimer par l'échantillonnage à deux phases), $Y = \sum_{i=1}^N y_i$.

x_{1i} = valeur de la covariable pour l'arbre i à la phase 1, par exemple le diamètre de l'arbre à hauteur de poitrine élevé au carré (D^2).

$X_1 = \sum_{i=1}^N x_{1i}$ (connu une fois la première phase appliquée à l'ensemble de la population).

$\pi_i(P)$ = probabilité de sélectionner l'arbre i dans l'échantillonnage de Poisson à une seule phase ($= n_{e1}x_{1i}/X_1$). Si tous les $\pi_i(P)$ sont égaux, il s'agit de l'échantillonnage binomial à une seule phase.

π_{1i} = probabilité de sélectionner l'arbre i à la phase 1 ($= n_{e1}x_{1i}/X_1$).

x_{2i} = valeur de la covariable pour l'arbre i à la phase 2, par exemple l'estimation visuelle du volume net.

X_2 = volume total estimé visuellement dans la population (uniquement obtenu pour les n_1 arbres sélectionnés à la première phase, de sorte que X_2 ne peut qu'être estimé).

π_{2i} = probabilité de sélectionner l'arbre i à la phase 2 ($= n_{e2}x_{2i}/\sum_{i=1}^{n_1} x_{2i}$).

y_i = valeur à l'étude pour l'arbre i (par exemple le volume net).

π_i = probabilité de sélectionner l'arbre i à l'application des deux phases de l'échantillonnage ($= \pi_{1i}\pi_{2i}$).

π_i^* = probabilité approximative de sélectionner l'arbre i suite à l'application des deux phases de l'échantillonnage ($= \pi_{1i}^*\pi_{2i}^*$ où $\pi_{1i}^* = n_{1i}x_{1i}/X_1$ et $\pi_{2i}^* = n_{2i}x_{2i}/\sum_{i=1}^{n_1} x_{2i}$).

Echantillonnage Poisson-Poisson et binomial-Poisson dans le domaine des forêts

Z. OUYANG, H.T. SCHREUDER, T. MAX et M. WILLIAMS¹

RÉSUMÉ

Les plans d'échantillonnage binomial-Poisson et Poisson-Poisson sont présentés en vue d'une utilisation dans le domaine des échantillonnages effectués en forêt. Plusieurs estimateurs du total de la population sont examinés pour ces plans. Des comparaisons (par simulation) des propriétés de ces estimateurs ont été faites pour trois petites populations forestières. Une modification de l'estimateur courant utilisé pour l'échantillonnage de Poisson, ainsi qu'un nouvel estimateur appelé estimateur de Srivastava modifié, semblent être les plus efficaces. Le dernier estimateur affiche malheureusement un biais prononcé pour les trois populations.

MOTS CLÉS: Peuplement forestier à valeur élevée; estimation de volume; estimateurs pour plan d'échantillonnage Poisson-Poisson; comparaisons par simulation; études portant sur les forêts; estimation de Srivastava.

1. INTRODUCTION

L'estimation des volumes, dans les études forestières, est un domaine très avancé, c'est-à-dire que l'on dispose de stratégies d'échantillonnage très efficaces pour estimer les volumes totaux (Schreuder et Ouyang 1992). Il est fréquent que l'estimation et la mesure du volume inexploitable (bois de rebut) ne soient pas prises en considération dans ces stratégies, parce que la mesure du volume inexploitable est difficile et qu'elle n'est pas justifiée, du point de vue économique, pour la plupart des peuplements. Toutefois, dans le cas des peuplements à valeur élevée, des stratégies à deux phases comme l'échantillonnage Poisson-Poisson, en vertu desquelles le volume inexploitable est mesuré sur les arbres à la deuxième phase, peuvent se révéler utiles. Pour l'étude des chargements de billes, l'échantillonnage binomial-Poisson peut constituer une technique convenable. Le but du présent article est de présenter la théorie des plans d'échantillonnage binomial-Poisson et Poisson-Poisson et d'examiner, par des simulations, certaines des propriétés d'estimateurs relatifs à ces plans.

2. REVUE DES TRAVAUX ANTÉRIEURS

Singh et Singh (1965) ont élaboré la théorie de l'échantillonnage à deux phases, avec probabilité proportionnelle à la taille (ppt) à la deuxième phase. Par ailleurs, Särndal et Swensson (1987) ont énoncé la théorie générale de l'échantillonnage à deux phases. Une liste d'unités d'échantillonnage est supposée disponible à la première phase, avant l'échantillonnage. Hajek (1957) a élaboré l'échantillonnage de Poisson, tandis que Groesenbaugh (1964) a proposé son application

à l'échantillonnage à une phase avec probabilités inégales dans les cas où aucune liste n'est disponible. L'échantillonnage de Poisson est une méthode selon laquelle chaque unité d'une population, disons l'unité i , est incluse dans l'échantillon de façon indépendante avec une probabilité p_i . Ainsi, la probabilité d'inclusion de l'unité i est égale à p_i , et la probabilité d'inclusion conjointe des unités i et j est égale à $p_i p_j$. L'échantillonnage binomial, souvent appelé aussi échantillonnage de Bernoulli, est un cas spécial de l'échantillonnage de Poisson dans lequel tous les p_i sont égaux. Dans les études portant sur les forêts, l'échantillonnage de Poisson comporte souvent les étapes suivantes (Schreuder et coll. 1968).

1. Visiter les N unités (par exemple des arbres) de la population dans n'importe quel ordre et mesurer ou estimer visuellement la valeur d'une covariable x_i ($i = 1, \dots, N$) en forte corrélation avec la valeur à l'étude y_i ($i = 1, \dots, N$).
2. Au moment de l'observation de chaque x_i , le comparer avec un entier δ_i pris au hasard dans l'intervalle $1 \leq \delta_i \leq L$, où L est un entier choisi avant l'échantillonnage. L est sélectionné de telle façon que $L = X/n_g$ où X est égal au total pour la covariable dans la population et n_g est la taille d'échantillon désirée. En général, X n'est pas connu avant l'échantillonnage et doit être estimé.
3. Si $\delta_i \leq x_i$, prendre l'unité dans l'échantillon et mesurer y_i .

L'application de cette méthode donne un échantillon de taille n , où $E(n) = n_g$ (si une bonne estimation de X a été effectuée avant l'échantillonnage). Dans l'échantillonnage binomial, tous les x_i ($i = 1, \dots, N$) sont les mêmes (Goodman 1949).

¹ Z. Ouyang, Anciennement boursier de recherches post-doctorales, Statistics Dept., Colorado State University, Fort Collins, Colorado, maintenant statisticien chercheur, ICI Seeds, Inc., Slater, Iowa; H.T. Schreuder, Chef de projet et statisticien, Multiresource Inventory Techniques Project, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado; T. Max, Biostatisticien, USDA Forest Service, Pacific Northwest Experiment Station, Portland, Oregon; M. Williams, Chef de projet et statisticien, Multiresource Inventory Techniques Project, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado; T. Max, Biostatisticien, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado.

BIBLIOGRAPHIE

- BRICK, J.M., et WAKSBERG, J. (1991). Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire. *Technique d'enquêtes*, 17, 31-46.
- BRUNNER, J.A., et BRUNNER, G.A. (1971). Are voluntarily unlisted telephone subscribers really different? *Journal of Marketing Research*, 8, 121-124.
- BURKHHEIMER, G.J., et LEVINSON, J.R. (1988). Implementing the Mitofsky-Waksberg sampling design with accelerated sequential replacement. Dans *Telephone Survey Methodology*, (Eds. R. Groves, et al.) 99-112. New York: John Wiley and Sons.
- GROVES, R.M. (1977). An Empirical Comparison of Two Telephone Designs. Rapport non-publié du Survey Research Center of the University of Michigan, Ann Arbor, MI.
- GROVES, R.M., et LEPKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 340-345.
- LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. Dans *Telephone Survey Methodology*, (Eds. R. Groves, et al.) 73-98. New York: John Wiley and Sons.
- MITOFSKY, W. (1970). Sampling of telephone households. Note de service inédite, CBS News, 1970.
- POTTHOFF, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.
- STOCK, J.S. (1962). How to improve samples based on telephone listings. *Journal of Advertising Research*, 2, 55-51.
- SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.
- SURVEY SAMPLING, INC. (1986). Statistical characteristics of random digit telephone samples produced by Survey Sampling, Inc. Westport, CT: Survey Sampling, Inc.
- TUCKER, C., CASADY, R.J., et LEPKOWSKI, J.M. (1992). Sample allocation for stratified telephone sample designs. A comparative, *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

La deuxième question afférente au coût touche les taux de succès plus faibles indiqués dans le présent document. Etant donné l'efficacité concurrentielle relative des méthodes de rechange envisagées ici, il semble que les taux inférieurs de succès ne constituent pas une entrave grave par rapport à l'efficacité des autres méthodes. On pourrait peut-être améliorer les taux de succès dans la strate à haute densité en recourant à des banques de numéros de plus petite taille. Par exemple, une autre étude démontre que les banques de 10 s'associent à des taux de succès d'environ 0,57 par rapport à celui de 0,52 indiqué ici pour les banques de 100. Evidemment, recourir à des banques de 10 accroît sensiblement la taille des dossiers et alourdit les opérations de traitement liées à la production des échantillons; le coût d'une base de banques de 10 est donc probablement bien supérieur à celui d'une base de banques de 100.

Les modèles de coûts indiqués en (2.2) et (2.3) sont relativement simples, puisqu'ils ne tiennent pas compte de nombreuses différences de coût du processus d'enquête téléphonique qui peuvent se révéler importantes dans la comparaison des efficacités relatives des plans d'échantillonnage. Ces modèles de coûts permettent l'expression simple de la répartition, mais ne traitent pas de façon spécifique des composantes de coût associées à deux aspects de la technique de Mitofsky-Waksberg, comme le font les autres plans d'échantillonnage, à savoir le remplissage des numéros inactifs et la pondération pour compenser l'épuisement des grappes. Par conséquent, les modèles de coûts n'intègrent pas certaines différences structurelles entre la technique Mitofsky-Waksberg et les autres options proposées, différences qui, le cas échéant, pourraient influencer sur les résultats de l'évaluation de l'efficacité relative des deux méthodes.

Il est clair qu'on ne peut tirer de conclusions définitives quant à la valeur globale de ces nouveaux plans à partir des seuls résultats décrits dans les présentes. Il faudra disposer de plus amples données sur les coûts ainsi que de données empiriques sur l'ampleur du biais qu'entraîne l'élimination des numéros de la strate de faible densité avant de pouvoir ainsi conclure.

REMERCIEMENTS

Nous remercions Clyde Tucker et Bob Groves de leur soutien et de leur aide. Les conclusions et opinions du présent document sont celles des auteurs et ne reflètent pas nécessairement celles du Bureau of Labor Statistics des Etats-Unis ni de l'Université du Michigan.

La thèse fondamentale du présent exposé veut que les méthodes d'échantillonnage stratifié, recourant à des strates basées sur le dénombrement de numéros de téléphone inscrits, sont au moins aussi efficaces que la technique de Mitofsky-Waksberg. De plus, ces plans d'échantillonnage peuvent éliminer la tâche fastidieuse de remplacement des numéros non résidentiels au deuxième degré, puisque les seuls numéros de téléphone qu'il faudra composer dans la strate de haute densité sont ceux qui auront été produits au début de l'enquête. Les conclusions spécifiques à tirer comprennent notamment les suivantes :

- dans le cas de faibles ratios des coûts, les plans à deux et à trois strates sont aussi efficaces que la méthode de Mitofsky-Waksberg;
- lorsqu'on peut laisser tomber les numéros de la strate de faible densité, ces autres plans d'échantillonnage sont beaucoup plus efficaces, mais au prix d'un biais inconnu qui découle de l'exclusion d'une partie de la population cible;
- dans le cas de ratios des coûts élevés, les méthodes à deux et à trois strates sont de toute évidence supérieures.

L'ampleur du biais qui découle de l'exclusion de la strate de faible densité constitue ici un élément critique. Comme il a été noté auparavant, environ 7% des ménages aux États-Unis ont pas le téléphone et il se peut que la concurrence de la base ajoute encore au biais lié à la non-trouver. Étant donné que la strate de faible densité contiendra vraisemblablement moins de 5% de la population des ménages aux États-Unis, il est probable que l'augmentation du biais de non-couverture ne sera pas substantielle à l'égard de beaucoup de caractéristiques de la population totale. Par ailleurs, relativement à certaines caractéristiques, ainsi qu'à certains sous-groupes de la population, l'importance du biais additionnel peut être suffisante pour justifier qu'on s'en préoccupe. Il faut alors procéder à d'autres analyses empiriques de cette population.

Il existe deux coûts afférents aux plans d'échantillonnage stratifié qui peuvent militer contre leur application : le coût de la liste commerciale utilisée pour stratifier la base BCR et le taux de succès global plus faible. Le coût de stratification de la base en strates de densité élevée et de faible densité n'est pas discuté ici puisque l'information requise est tirée d'un dossier de recherche spécialisée. Le coût de stratification est un coût fixe et entrainera par conséquent les ressources disponibles pour la collecte des données. On ne connaît pas la taille future de ce coût fixe puisqu'elle dépend d'ententes périodiques avec les entreprises commerciales qui procurent ces données. De plus, ce coût fixe de stratification peut s'amortir par rapport à plusieurs études, ce qui en réduit grandement les répercussions sur chacun des échantillons. Il est improbable que la collecte des données pour une seule enquête soit aussi efficace au plan des coûts qu'il ne l'est indiqué, que ce soit dans le cas de la méthode de Mitofsky-Waksberg ou dans celui de la méthode de stratification décrite dans les présentes. Il faudra analyser plus à fond les coûts liés à la base avant de disposer d'une réponse complète à cet égard.

sorte que ces résultats peuvent être considérés d'application générale en présence de variables analytiques de type Bernoulli. Dans tous les cas, la réduction réelle de la variance est pratiquement identique à celle réalisée dans des conditions de répartition optimale; par conséquent, la répartition selon la formule (2.6) peut être tenue pour (presque) optimale. La réduction prévue de la variance s'est également révélée très proche de la réduction réelle. Lorsque p_2 était inférieur à p_1 , la réduction réelle était toujours supérieure à la réduction prévue, et l'inverse s'avérait lorsque p_2 était supérieur à p_1 . Dans les deux cas, la différence maximale (qui n'était que d'environ 3,5% de la réduction réelle pour $\gamma = 2$ et de 8,3% pour $\gamma = 10$) est survenue pour $p_1 = 0.05$ et décroissait monotonicement avec la croissance de p_1 .

En résumé, les deux cas à l'étude semblent indiquer que tant que les hypothèses qui donnent la répartition obtenue par l'application de la formule (2.6) ne sont pas carrément violées, la variance sera très proche de celle que l'on obtient dans des conditions de répartition optimale. De plus, la réduction proportionnelle de la variance obtenue selon la formule (2.7) constitue une estimation de la réduction réelle de la variance au moins suffisamment précise pour satisfaire les besoins des plans d'enquête.

5. CONCLUSIONS

Les points forts de la technique de Mitofsky-Waksberg en matière de production d'échantillons téléphoniques sont évidents : des taux de succès élevés dans la sélection de deuxième degré, une méthode efficace de tri des banques vides de numéros de téléphone, et un concept ingénieux pour aborder la production d'échantillon. Le fait que cette technique soit largement considérée comme la méthode type de composition aléatoire, et qu'on ne lui reconnaisse que peu de méthodes concurrentes après de nombreuses années d'existence, témoigne éloquentement de sa force. Ses points faibles (le tri de premier degré et le remplacement des numéros non résidentiels lors de la collecte des données) ne semblent pas, à prime abord, importer par rapport à sa force globale. Toutefois, ces caractéristiques peuvent soulever des difficultés importantes, particulièrement lors d'enquêtes téléphoniques soumises à une contrainte de courte durée.

Le présent document propose, comme méthodes de rechange à celle de Mitofsky-Waksberg, des plans d'échantillonnage stratifié, basés sur des listes commerciales de numéros de téléphone. Les plans à deux et à trois strates y sont étudiés dans le détail. En plus de l'échantillonnage aléatoire simple dans chaque strate, deux autres options générales y sont considérées :

- (1) l'échantillonnage aléatoire simple sur toutes les strates sauf celle de faible densité pour laquelle on applique la méthode de Mitofsky-Waksberg, et
- (2) l'échantillonnage aléatoire simple sur toutes les strates, sauf celle de faible densité qui ne fait l'objet d'aucun échantillonnage.

ces questions en ce qui touche le plan d'échantillonnage à deux strates relativement à trois cas précis d'échec du plan, dans des circonstances typiques du "monde réel". Dans ces trois cas, les résultats indiquent une réponse nettement affirmative aux deux questions.

Dans le premier cas, supposons que $\sigma_1^2 = \sigma_2^2 = W^2$ mais que $\lambda_1 \neq \lambda_2$. La réduction prévue, la réduction réelle, et la réduction maximale de la variance ont été calculées pour diverses valeurs données de $\beta = |\sqrt{\lambda_1} - \sqrt{\lambda_2}| = |\mu_1 - \mu_2|/W$ entre 0.00 et 0.50; les résultats figurent au Tableau 5 ci-dessous. Selon nos discussions précédentes au sujet de la faible relation entre les variables analytiques et les variables qui servent à la stratification, il semble hautement improbable que β puisse dépasser 0.50. Les résultats qui figurent au Tableau 5 indiquent que dans le cas des deux ratios des coûts pour toutes les valeurs choisies de β , la réduction réelle de la variance obtenue par la répartition d'échantillonnage en appliquant les hypothèses simplificatrices est fondamentalement équivalente à celle que l'on obtiendrait dans des conditions de répartition "optimales". Pour les deux ratios des coûts, la réduction prévue de la variance est toujours supérieure à la réduction réalisée de fait, l'écart augmentant à mesure que β s'accroît. Toutefois, on devrait prendre note que pour $\beta \leq .35$, la différence en pourcentage entre la réduction prévue et la réduction réelle est inférieure à 10 % lorsque $\gamma = 10$, et inférieure à 4 % lorsque $\gamma = 2$.

Tableau 5

La réduction proportionnelle prévue, actuelle et maximale de la variance pour les ratios des coûts de 2 et de 10 et les valeurs de β entre 0.00 et 0.50

β	$\gamma = 2$					$\gamma = 10$				
	Réduction	Réduction	Réduction	Réduction	Réduction	Réduction	Réduction	Réduction	Réduction	Réduction
	proje- tion réelle	proje- tion maxi- male	proje- tion réelle	proje- tion maxi- male	proje- tion réelle	proje- tion réelle	proje- tion maxi- male	proje- tion réelle	proje- tion maxi- male	proje- tion maxi- male
0.00	.2829	.2829	.2829	.2829	.2829	.0766	.0766	.0766	.0766	.0766
0.10	.2829	.2820	.2820	.2820	.2820	.0766	.0766	.0761	.0761	.0761
0.20	.2829	.2793	.2794	.2794	.2794	.0766	.0766	.0745	.0745	.0746
0.30	.2829	.2748	.2750	.2750	.2750	.0766	.0766	.0720	.0720	.0721
0.40	.2829	.2686	.2692	.2692	.2692	.0766	.0766	.0684	.0684	.0689
0.50	.2829	.2607	.2619	.2619	.2619	.0766	.0766	.0639	.0639	.0649

Dans un deuxième cas d'application générale, il est sup-
posé que la variable analytique est celle de Bernoulli, où p_1 et p_2 représentent la proportion de la population dotée de l'attribut d'intérêt dans la strate 1 et la strate 2 respec-
tivement. La réduction proportionnelle prévue, la réduction
réelle et la réduction maximale de la variance a été calculée
pour deux cas précis où l'hypothèse est infirmée, à savoir
 $p_2 = .90p_1$ et $p_2 = 1.10p_1$; on a fait varier p_1 de .05 à .50
et considéré des ratios des coûts de 2 et 10.
Comme nous l'avons vu auparavant, il est probablement
raisonnable de supposer que p_2 demeurera à moins de 10%
près de p_1 dans la plupart des situations "réelles" de telle

nombre de NR de certaines banques. Ces deux difficultés
sont évidemment plus marquées lorsque $\gamma = 2$ que
lorsque $\gamma = 10$. En conséquence, le recours à ce plan
d'échantillonnage est réservé aux situations où les res-
sources peuvent permettre un échantillon de "grande"

taille, et le ratio des coûts est de modéré à élevé.

Tableau 4

Ratios de répartition au premier degré et tailles
de l'échantillon de deuxième degré pour
le plan d'échantillonnage combiné composition
aléatoire/Mitofsky-Waksberg appliqué
à la base BCR à deux strates

Strate	$\gamma = 2$		$\gamma = 10$	
	Taille de l'échantillon m_1/\bar{m}_2	degré	Taille de l'échantillon m_1/\bar{m}_2	degré
1	28.17	S.O.	14.56	S.O.
2	S.O.	17.00	S.O.	9.00

4. RÉPARTITION D'ÉCHANTILLON ET EFFICACITÉ DU PLAN

La section 2.6 traitait du problème de la détermination
des paramètres nécessaires à la répartition optimale de
l'échantillon dans l'ensemble des strates. Il y était noté que
les paramètres spécifiques des variables (c.-à-d., λ_i et σ_i^2)
sont ceux qui ont tendance à présenter la difficulté princi-
pale puisque que l'on dispose en général de peu d'infor-
mation quant à leurs valeurs. Dans la plupart des cas, les
variables d'intérêt analytique ne sont pas très étroitement
liées aux variables qui servent à la stratification. Il est par
conséquent raisonnable de supposer que $\lambda_i = 0$ et que
 $\sigma_i^2 = \sigma^2$ pour $i = 1, 2, \dots, H$. Étant donné ces hypothèses,
la répartition optimale s'obtient par la formule (2.6) et la
réduction proportionnelle de la variance s'obtient par la
formule (2.7).
Il est évident que ces hypothèses ne s'appliquent jamais
parfaitement lors d'une application particulière, et que
lorsque l'on détermine la répartition au moyen de l'équa-
tion (2.6) la réduction proportionnelle réelle de la variance
ne correspond pas exactement au résultat du calcul de
l'équation (2.7). De plus, une répartition conforme aux
données de (2.6) ne rendra pas la réduction maximale de
la variance que l'on obtient dans les conditions de répar-
tition optimale conformes aux données de (2.4). Supposons
que nous voulons répartir l'échantillon en conformité de
la formule (2.6); deux questions doivent être tranchées:
la formule (2.7) procure-t-elle une approximation raison-
nable de la réduction réelle de la variance? et (2), la réduction
réelle de la variance est-elle raisonnablement proche de la
réduction maximale possible? Il n'existe aucune réponse
unique simple à ces questions étant donné que le résultat
dépend de la manière et de l'ampleur exactes selon laquelle
les hypothèses s'infirmement. Nous discuterons ci-après de

L'estimateur de ratio combiné $\bar{Y}_s = \bar{Y}_s / \bar{N}_s$, où $\bar{Y}_s = \sum_{i=1}^H M_i / m_i \bar{y}_i + \sum_{i=H+1}^H M_i / \bar{m}_i (\bar{y}_i / k_i + 1)$ et $\bar{N}_s = \sum_{i=1}^H M_i / m_i \bar{m}_i + \sum_{i=H+1}^H M_i / \bar{m}_i \bar{m}_i$, sert à estimer la moyenne μ de la population et les valeurs de m_i et k_i doivent être choisies pour minimiser la var (\bar{Y}_s) prévue, ou le coût, selon le cas.

L'estimateur \bar{Y}_s est non biaisé asymptotiquement pour μ et on peut démontrer directement que

$$\text{var}(\bar{Y}_s) \equiv \sum_{i=1}^H \frac{z_i^2 \sigma_i^2}{m_i h_i} (1 + (1 - h_i) \lambda_i)$$

$$+ \sum_{i=H+1}^H \frac{z_i^2 \sigma_i^2}{m_i h_i} [1 + (1 - h_i) \lambda_i]$$

$$- k_i (1 - \rho) (k_i + 1)^{-1}] \quad (3.4)$$

et

$$E[C(D_s)] = c_0 \left\{ \sum_{i=1}^H m_i [1 + h_i (\gamma - 1)] \right.$$

$$+ \sum_{i=H+1}^H m_i [1 + k_i (1 - t_i)]$$

$$\left. + h_i (k_i + 1) (\gamma - 1) \right\}. \quad (3.5)$$

Les valeurs optimales de m_i et \bar{m}_i , spécifiées à une constante de proportionnalité près s'obtiennent par la formule suivante

$$m_i \propto z_i \sigma_i \left(\frac{1 + (1 - h_i) \lambda_i}{h_i (1 + h_i (\gamma - 1))} \right)^{1/2}, \quad (3.6)$$

pour $i = 1, \dots, H$ et

$$\bar{m}_i \propto z_i \sigma_i \left(\frac{\lambda_i (1 - h_i) + \rho}{h_i t_i} \right)^{1/2}, \quad (3.7)$$

pour $i = H_1 + 1, \dots, H$. La valeur optimale de $(k_i + 1)$, pour $i = H_1 + 1, \dots, H$, s'obtient par la formule suivante

$$k_i + 1 = \max$$

$$\left\{ 1, \left(\frac{t_i (1 - \rho)}{(1 + h_i (\gamma - 1) - t_i) (\lambda_i (1 - h_i) + \rho)} \right)^{1/2} \right\}. \quad (3.8)$$

La constante de proportionnalité pour (3.6) et (3.7), s'obtient par substitution dans l'équation du coût prévu ou l'équation de la variance selon le cas.

Dans des conditions de répartition optimale, la réduction de la variance (ou du coût) par rapport à la composition aléatoire simple, s'obtient par la formule suivante

$$R(\bar{Y}_s, \bar{Y}_0) = 1 - \frac{\bar{h} \Phi^2}{\sigma^2 (1 + (\gamma - 1) \bar{h})}, \quad (3.9)$$

où

$$\Phi = \sum_{i=1}^H \frac{z_i^2 \sigma_i^2}{h_i} (1 + (1 - h_i) \lambda_i)^{1/2} (1 + (\gamma - 1) h_i)^{1/2}$$

$$+ \sum_{i=H+1}^H \frac{z_i^2 \sigma_i^2}{h_i} \left[\rho + (1 - h_i) \lambda_i \right]^{1/2} t_i^{1/2} +$$

$$(1 - t_i + (\gamma - 1) h_i)^{1/2} (1 - \rho)^{1/2}. \quad (3.10)$$

Avec l'hypothèse simplificatrice, $\lambda_i = 0$ et $\sigma_i^2 = \sigma^2$ pour $i = 1, 2, \dots, H$,

$$\Phi = \sigma \left[\sum_{i=1}^H \frac{z_i^2 h_i^{1/2}}{h_i} (1 + (\gamma - 1) h_i)^{1/2} \right]$$

$$+ \sigma \left[\sum_{i=H+1}^H \frac{z_i^2 h_i^{1/2}}{h_i} (\rho t_i)^{1/2} + \right.$$

$$\left. \left. (1 - t_i + (\gamma - 1) h_i) (1 - \rho) \right)^{1/2} \right]. \quad (3.11)$$

Lorsqu'elle est appliquée à une base à deux strates, cette stratégie d'échantillonnage combiné procure une réduction proportionnelle de la variance d'environ $R = .440$ pour $\gamma = 2$ et de $R = .157$ pour $\gamma = 10$. Pour les deux ratios des coûts, la réduction de la variance est sensiblement supérieure à celle obtenue par l'application de toute méthode non biaisée discutée auparavant. En fait, la réduction de la variance est essentiellement équivalente à celle que l'on obtient en appliquant le plan d'échantillonnage tronqué à trois strates (qui est sujet à un biais d'ampleur inconnu). Par conséquent, à première vue, cette stratégie d'échantillonnage combiné semble supérieure à toutes les autres méthodes.

Malheureusement, certaines contraintes d'ordre pratique peuvent interdire le recours à ce plan d'échantillonnage dans certaines situations. Par exemple, le taux de succès dans la strate de Mitofsky-Waksberg est très bas (seulement .02), de sorte que le nombre de banques de 100 dans l'échantillon de premier degré doit être passablement élevé afin que le nombre prévu de banques de 100 retenues ne soit pas trop faible. D'autre part, le nombre *relatif* d'unités de l'échantillon de premier degré réparties à la strate de celui de la strate de Mitofsky-Waksberg et l'échantillon global doit donc être de grande taille (voir le tableau 4). De même, comme l'indique également le tableau 4, le nombre de NR requis de chacune des banques de 100 retenues est relativement élevé et peut même dépasser le

(Bien qu'il n'en soit pas discuté ici, le Tableau 3 indique également la réduction prévue de la variance pour un ratio des coûts de 20.)

Tableau 3

Réduction proportionnelle prévue de la variance (ou des coûts), par rapport à l'échantillonnage par composition aléatoire simple, pour cinq différents plans d'échantillonnage téléphonique

Plan d'échantillonnage	Proportion de la base exclue du champ d'observation			Proportion de la variance ou du coût	γ = 2	γ = 10	γ = 20
	Réduction proportionnelle	de la variance	exclue du champ d'observation				
Deux strates	.2829	.0766	.0320	.0598	.4917	.2055	.1189
Deux strates (tronquée)				.0000	.2811	.0597	.0135
Mitofsky-Waksberg				.0000	.3001	.0866	.0389
Trois strates				.0000	.4095	.1574	.0879
Trois strates (tronquée)				.0199			

La stratégie de répartition proposée réussit à réduire le pourcentage de la population exclue du champ d'observation, de presque de 6% à environ 2%. La réduction proportionnelle prévue de la variance pour le plan à trois strates tronqué est d'environ $R = .410$ lorsque $\gamma = 2$ et $R = .157$ lorsque $\gamma = 10$. Du point de vue de l'efficacité, ce modèle se situe entre le modèle hautement efficace à deux strates tronqué et les modèles non biaisés. La préoccupation principale que soulève la perspective du choix d'un tel modèle est bien sûr liée au problème de couverture. Le plan d'échantillonnage est déjà sujet à la non-couverture de la population des ménages qui n'ont pas le téléphone. Tronquer la base peut ajouter au biais de non-couverture qui découle déjà de cette cause. Aux fins de toute application spécifique, on doit évaluer le risque inhérent à l'échantillonnage tiré d'une base qui ne compte pas toutes les unités de la population cible en regard du gain possible d'efficacité. Comme on peut le prévoir, le modèle type à trois strates est légèrement plus efficace que celui à deux strates. Toutefois, l'augmentation d'efficacité est tellement faible qu'on peut douter du bien-fondé du coût additionnel de la répartition de la base BCR pour ajouter une strate, si ce n'est dans le but de procéder à la troncature.

3.2 Plans d'échantillonnage incluant la répartition optimale et la méthode de Mitofsky-Waksberg

Le dernier plan à l'étude est fondé sur la base BCR stratifiée. Il recourt à l'échantillonnage à composition aléatoire simple dans certaines strates ou à l'échantillonnage

de Mitofsky-Waksberg dans d'autres strates, selon la proportion de banques de 100 vides dans la strate. Les deux considérations suivantes justifient le choix de ce type de plan ou modèle d'échantillonnage:

a) l'échantillonnage de Mitofsky-Waksberg a tendance à être "complexe au plan administratif", et lorsque le gain d'efficacité est faible, on lui préfère l'échantillonnage à composition aléatoire simple;

b) il découle de l'équation (2.9), appliquée au niveau de la strate, que lorsque la proportion des banques vides dans une strate est "faible", l'échantillonnage de Mitofsky-Waksberg offre peu ou pas d'augmentation d'efficacité.

Par conséquent, nous proposons de recourir à l'échantillonnage par composition aléatoire simple dans les strates à "faible" proportion de banques de 100 vides et de recourir à l'échantillonnage de Mitofsky-Waksberg dans les autres strates. Le choix du type d'échantillonnage des autres strates. Plus précisément, si le nombre total "optimal" de NR, tel que le détermine l'équation (2.8), à choisir d'un échantillon de banques de 100 dans une strate donnée est égal à un, alors la strate est désignée une strate à composition aléatoire simple; dans les autres cas, elle est désignée une strate Mitofsky-Waksberg. En termes de la proportion des banques de 100 vides, la i^{e} strate sera une strate à composition aléatoire si

$$t_i \leq \frac{2.25p(1 + h_i(\gamma - 1))}{(1 + 1.25p)} \quad (3.3)$$

et, sinon, une strate Mitofsky-Waksberg. Dans l'exemple à deux strates, la première strate est une strate à composition aléatoire et la seconde, une strate Mitofsky-Waksberg lorsque γ égale 2 ou 10. Le plan d'échantillonnage proposé peut se définir formellement ainsi qu'il suit. La base BCR a été répartie en H strates et, en fonction du critère donné en (3.3), l'échantillonnage à composition aléatoire simple est choisi pour les H_1 premières strates alors que l'échantillonnage Mitofsky-Waksberg est retenu pour les autres strates.

Soient:

m_i = le nombre de numéros de téléphone choisis de la i^{e} strate à composition aléatoire,

m'_i = le nombre de banques de 100 choisies de la i^{e} strate Mitofsky-Waksberg,

m''_i = le nombre de banques de 100 retenu de la i^{e} strate Mitofsky-Waksberg,

k_i = le nombre de NR additionnels choisis de chaque banque de 100 retenue, et

y_i = agrégat des valeurs y pour les NR de l'échantillon tirés de la i^{e} strate.

3. AUTRES PLANS D'ÉCHANTILLONNAGE

3.1 Plans tronqués

Les plans d'échantillonnage présentés à la section précédente produisaient des estimés non biaisés de la moyenne de la population. Une hypothèse incorrecte au sujet des divers paramètres de la base, des coûts et de la population n'affecte que l'efficacité des estimateurs, et non leurs prévisions. Malheureusement, il faut payer un prix très élevé pour s'assurer de l'absence de biais, l'échantillonnage sur la strate résiduelle ne procurant de l'information qu'au sujet d'une proportion minime de la population, et ce à un coût relativement élevé. Par exemple, supposons que l'on accepte de se contenter d'une estimation de la moyenne de la population à l'exclusion des ménages associés aux numéros de la strate résiduelle (c.-à-d., que nous "tronquons" la base initiale en éliminant la strate résiduelle et tirons un échantillon stratifié par composition aléatoire à partir des autres numéros de téléphone). Dans le cas de l'exemple à deux strates, la "base tronquée" ne comprendrait que les numéros de téléphone de la première strate. Le taux de succès de l'échantillon de la base tronquée serait de .521, par opposition à un taux de succès de .211 pour la base complète. Toutefois, le champ d'observation ne comprendrait que seulement 94% environ de la population cible.

Dans les lignes suivantes, nous supposons que la base tronquée est constituée de la simple base BCR initiale moins la strate résiduelle que nous supposons (sans perte de généralité) être la strate H . En conséquence, pour la base tronquée $h^* = (h - P^H h^H) / (1 - P^H)$ est le taux de succès, $t^* = (t - P^{Ht} t^H) / (1 - P^H)$ est la proportion des banques de 100 vides et $\mu^* = (\mu - z^H \mu^H) / (1 - z^H)$ est la moyenne de cette population. Le plan D_4 étant l'échantillonnage aléatoire simple stratifié de la base tronquée, et \bar{Y}_4 l'estimateur type du ratio (du quotient) de la moyenne de la population. L'estimateur \bar{Y}_4 est non biaisé asymptotiquement pour μ^* , et, en général, biaisé pour μ . Le biais (asymptotique) s'obtient par la formule

$$B(\bar{Y}_4) = \mu^* - \mu = \frac{z^H(\mu - \mu^H)}{(1 - z^H)} \quad (3.1)$$

Dans la plupart des cas pratiques, le biais tend monotoniquement vers zéro lorsque la proportion de la population cible de la strate résiduelle devient faible, même si, comme l'indique (3.1), ce n'est pas nécessairement toujours le cas. De toute façon, puisque la valeur de $\mu - \mu^H$ n'est jamais connue, la valeur limite supérieure à fixer à la proportion de la population de la strate résiduelle constitue habituellement la spécification clé qu'il faut déterminer lorsqu'on envisage le recours à une base tronquée. Dans l'exemple à deux strates, environ 6% de la population cible est exclue de la base d'échantillonnage ce qui, dans presque tous les cas, serait inacceptable à un organisme fédéral. Les formules du coût, de la variance, de la répartition et de la réduction proportionnelle de la variance (ou du coût) sont essentiellement les mêmes que celles qui figurent

à la section n° 2. De fait, les seules modifications à apporter à la formule (2.1) et aux formules (2.3) à (2.7) consistent à remplacer μ par μ^* et, pour $i = 1, 2, \dots, H - 1$, à remplacer z_i par $z_i^* = z_i / (1 - z^H)$ et à remplacer λ_i par $\lambda_i^* = (\mu_i - \mu^*)^2 / \sigma_i^2$. Évidemment, toutes les sommes ne valent que pour la strate restante $H - 1$. Dans le cas spécial où il ne reste qu'une strate après la troncature, la réduction proportionnelle de la variance (du coût) se réduit à la formule suivante

$$R(\bar{Y}_4, \bar{Y}_0) = 1 - \frac{h^*(1 + h^*(\gamma - 1))}{h(1 + h(\gamma - 1))} \quad (3.2)$$

Afin de tenter de maintenir l'efficacité relative de la troncature tout en diminuant la partie du problème de couverture, le BLS et l'Université de Michigan étudient plusieurs autres plans de stratification qui visent à réduire la proportion de la population de la strate résiduelle. Une avenue prometteuse s'appuie sur la répartition de la strate résiduelle en deux strates résiduelles ou plus. Par exemple, la répartition pourrait produire une strate résiduelle n° 3 comprenant les numéros de téléphone de la banque de 100 presumed attribues en premier lieu à des établissements commerciaux ou non encore actives à des fins résidentielles ou commerciales. La strate résiduelle n° 2 comprendra alors tous les autres numéros de téléphone de la strate résiduelle obtenue avec le plan à deux strates D_2 . Les paramètres estimés de la base pour le modèle à trois strates qui en découle paraissent au Tableau 2.

Tableau 2

Paramètres de la base estimés d'un projet de plan à trois strates fondés sur la base BCR et sur la base de la liste Donnelley

Strate	Proportion de la population	Proportion de succès	Taux de banques vides	Taux de succès dans les banques non vides
1	.3804	.9402	.5210	.0300
2	.2000	.0399	.0420	.9143
3	.4196	.0199	.0100	.9796
				.4900

Les données ci-dessus ont servi à calculer la réduction proportionnelle prévue de la variance pour un plan à trois strates et un plan à trois strates tronqué dont la strate n° 3 est exclue. Les résultats, de même qu'un sommaire des résultats des plans à deux strates et du plan de Mitofsky-Waksberg figurent au Tableau 3 ci-dessous.

$$R(\bar{Y}_1, \bar{Y}_0) \equiv 1 -$$

$$\left[\sum_{i=1}^I \frac{z_i \sigma_i}{h_i} \right]_{\frac{1}{2}} \frac{h^{-1} \sigma^2 (1 + (\gamma - 1)h)}{2} \quad (2.5)$$

2.6 Problèmes d'ordre pratique associés à la répartition optimale

Le problème de la détermination des valeurs des paramètres des équations de répartition est un problème général à tous les scénarios de répartition optimale. Dans le cas qui nous intéresse, trois types fondamentaux de paramètres entrent en ligne de compte: les paramètres liés à la base (z_i et h_i), ceux reliés au coût (γ et c_0) et les paramètres spécifiques à la variable d'intérêt (λ_i et σ_i^2). Nous disposons de nos jours d'une assez bonne connaissance pratique des paramètres liés à l'exemple à deux strates et à certains autres modèles spécifiques de stratification. Nous discuterons, à la section n° 5, de plusieurs projets de recherche présentant en cours qui devraient permettre d'approfondir les connaissances en ce domaine.

Il est clair que $\gamma \geq 1$, mais la valeur réelle peut varier sensiblement. Par exemple, dans le cas d'une enquête polyvalente, on rassemble l'information pour plusieurs variables, et les coûts de détermination de la nature résidentielle ou non d'un numéro de téléphone, c_0 et c_1 , sont dans les faits amortis sur les variables d'intérêt, et γ sera probablement considérablement supérieur à l'unité. Par ailleurs, si l'enquête ne vise à rassembler des renseignements que sur une variable unique, la valeur de γ n'est alors probablement guère supérieure à deux ou trois. Waksberg (1978) considère

des valeurs de γ entre 2 et 20. Potentiellement, les paramètres spécifiques aux variables présentent la plus grande difficulté. En général, notre connaissance des valeurs de ces paramètres est restreinte et, dans le cas des enquêtes polyvalentes, il nous faut décider quelle(s) variable(s) appliquer aux fins de la répartition. Heureusement, dans plusieurs cas pratiques, deux facteurs se combinent dans le sens d'une certaine réduction de cette difficulté. En premier lieu, la répartition a tendance à être relativement "stable" dans le voisinage de la répartition optimale de telle sorte que la réduction de la variance est relativement robuste en regard de la répartition. En second lieu, dans la plupart des cas, les variables d'intérêt ne sont pas étroitement liées aux variables de type qui sert à la stratification. Par conséquent, avec prudence, nous supposons que $\lambda_i = 0$ et $\sigma_i^2 = \sigma^2$ pour $i = 1, 2, \dots, H$. La répartition optimale se réalise dans les conditions suivantes

$$m_i \propto \frac{z_i}{\lambda_i} \frac{1 + (\gamma - 1)h_i}{h_i} \quad (2.6)$$

et la réduction proportionnelle de la variance est

$$R(\bar{Y}_1, \bar{Y}_0) \equiv 1 - h \frac{\left[\sum_{i=1}^I z_i \left(\frac{1 + (\gamma - 1)h_i}{h_i} \right) \right]_{\frac{1}{2}}}{(1 + (\gamma - 1)h)} \quad (2.7)$$

Dans le cas de l'exemple à deux strates, la répartition spécifiée en (2.6) implique que la répartition relative à la strate résiduelle (c.-à-d., m_1/m_2) est de 2,54 lorsque $\gamma = 2$ et 1,42 lorsque $\gamma = 10$. Dans le premier cas, la réduction proportionnelle projetée de la variance est $R = .283$ et, dans le second cas, $R = .077$. En fait, on peut déduire de (2.7) que lorsque le coût relatif de la détermination de la valeur de la variable d'intérêt augmente, l'avantage relatif de la répartition optimale décroît. Le plan d'échantillonnage de Mitofsky-Waksberg, désigné D_3 , applique une sélection d'échantillon à deux degrés (c.-à-d., des banques de 100 non vides sont choisies au premier degré et des NR sont choisis au deuxième degré). Selon Waksberg (1978), nous laissons $(k + 1)$ être le nombre total de NR choisis de chaque banque de 100 de l'échantillon. L'estimateur de Mitofsky-Waksberg, désigné \bar{Y}_3 , est non biaisé par rapport à μ , et sa variance est minimale lorsque

$$k + 1 = \max \left\{ 1, \left(\frac{(1 - p)\bar{t}}{(1 + (\gamma - 1)h - \bar{t})p} \right)^{\frac{1}{2}} \right\}, \quad (2.8)$$

où p est la corrélation intra-bancaire. Dans cette condition de répartition "optimale" au sein de l'échantillon de banque de 100, la réduction de la variance, relativement à une composition aléatoire simple, pour l'estimateur \bar{Y}_3 s'obtient par la formule

$$R(\bar{Y}_3, \bar{Y}_0) \equiv 1 - \frac{(1 + (\gamma - 1)h)}{\left[(1 + (\gamma - 1)h - \bar{t})^{\frac{1}{2}} + (p\bar{t})^{\frac{1}{2}} \right]^2} \quad (2.9)$$

Au niveau national, Groves (1977) rapporte que $p \approx .05$ dans le cas des statistiques économiques ou sociales. En se servant de cette valeur de p , ainsi que des valeurs de \bar{h} et \bar{t} de l'exemple à deux strates, la réduction proportionnelle prévue de la variance pour la méthode Mitofsky-Waksberg est $R = .281$ lorsque $\gamma = 2$ et $R = .060$ lorsque $\gamma = 10$. Les deux méthodologies semblent donner lieu à une réduction de la variance pratiquement identique pour les deux valeurs du ratio des coûts. Toutefois, il faut se garder de trop conclure à partir de cette simple comparaison, puisque la réduction prévue dans chacune des méthodes se fonde sur des hypothèses simplificatrices qui ne s'avèrent pas strictement dans toutes les applications. La seule hypothèse posée ici est que les deux méthodologies semblent hautement concurrentielles dans un ensemble général de circonstances habituellement présentes dans la pratique.

$$C_1(d) = \begin{cases} c_1 & \text{si } d = 1 \\ c_0 & \text{si } d = 0 \end{cases}$$

d'intérêt Y . La fonction des coûts pour déterminer la variable indicatrice est désignée par $C_1(\cdot)$, avec

Ce modèle prévoit la possibilité que le coût de détermination qu'un numéro de téléphone n'est pas un NR soit différent de celui de détermination qu'il en est un. En fait, le coût de détermination de la nature résidentielle ou non d'un numéro de téléphone est habituellement moindre s'il s'agit d'un NR. Le coût de détermination de la valeur de la caractéristique X comprend seulement le *coût additionnel* de la détermination de la valeur de Y une fois la valeur de d déterminée. Soit $C_2(\cdot, \cdot)$ qui représente ce coût additionnel, avec

$$C_2(d, y) = \begin{cases} 0 & \text{si } d = 0 \\ c_2 & \text{si } d = 1 \end{cases} \quad \text{si le } j^{\text{e}} \text{ numéro de téléphone de la } i^{\text{e}} \text{ strate est un NR,} \\ 0 & \text{sinon.}$$

Supposons que les numéros de téléphone de la i^{e} strate sont étiquetés 1 à M_i . Soit

2.3 Le problème de base de l'estimation, les plans d'échantillonnage et les estimateurs

prendre note que pour la base BCR, $h \approx .211$ et $\bar{t} \approx .605$. La valeur de h est très près de celle que donne Wakseberg (1978), mais la valeur de \bar{t} est légèrement inférieure à celle de .65 déterminée par Groves (1977). Il est présentement impossible de déterminer laquelle des valeurs de \bar{t} est la plus précise; en fait, cette valeur peut avoir changé depuis 1977. Plus récemment, Tucker, Casady et Lepkowski (1992) ont estimé la valeur de \bar{t} à .616 pour des banques de 10, ce qui tend à appuyer la moindre estimation de \bar{t} pour les banques de 100.

La variable d'intérêt est la caractéristique du ménage Y , et y représente la valeur de Y pour un ménage spécifique. Le paramètre de la population à estimer est la moyenne de la population $\mu = X/N$, où $N = \sum_{i=1}^H \sum_{j=1}^{M_i} d_{ij} y_{ij}$. Le terme N_i dénote le nombre de NR dans la i^{e} strate et N dénote le nombre de NR dans la population.

Considérons deux plans d'échantillonnage: (1) un échantillonnage aléatoire simple sans remplacement sur les numéros de téléphone de la base BCR, désigné plan D_0 et (2) un échantillonnage aléatoire simple stratifié sur la base BCR (c.-à-d., des échantillons indépendants sont choisis par échantillonnage aléatoire simple de chaque strate), désigné plan D_1 . Dans le plan D_0 , l'estimateur type du ratio pour μ s'obtient par la formule $X_0 = Y_0/N_0$ où X_0 et N_0 sont les estimateurs types d'extrapolation pour Y et N , respectivement. L'estimateur X_0 est sans biais asymptotique pour μ et sa variance s'obtient par la formule $\text{var}(X_0) \approx \sigma^2/mh$ où m est la taille de l'échantillon des numéros de téléphone et σ^2 est la variance de la population des y . Dans le plan d'échantillonnage D_1 , l'estimateur du ratio type pour μ s'obtient par la formule $X_1 = Y_1/N_1$ où Y_1 et N_1 sont les estimateurs types d'extrapolation pour Y et N , dans l'échantillonnage stratifié. L'estimateur X_1 est également sans biais asymptotique pour μ et sa variance s'obtient par la formule

$$\text{var}(X_1) \approx \sum_{i=1}^H \frac{z_i^2 \sigma_i^2 (1 + (1 - h_i) \lambda_i)}{m_i h_i}, \quad (2.1)$$

où $\lambda_i = (\mu_i - \mu)^2/\sigma_i^2$ et où m_i , μ_i , et σ_i^2 représentent respectivement les tailles, les moyennes, et les variances de l'échantillon des strates.

2.4 Le modèle de coûts

Des coûts sont associés à la détermination de la valeur de la variable indicatrice d et de la valeur de la caractéristique

dans laquelle la constante de proportionnalité est déterminée par substitution dans l'équation des coûts prévus (ou dans l'équation de la variance, selon le cas). La réduction proportionnelle de la variance, par rapport à l'échantillonnage par composition aléatoire, dans les conditions de répartition optimale pour un coût déterminé C^* (ou la réduction proportionnelle des coûts dans les conditions de répartition optimale pour une variance déterminée V^*) s'obtient par la formule suivante

$$m_i \propto \frac{z_i \sigma_i}{\sqrt{h_i}} \left(\frac{1 + (\gamma - 1) h_i}{1 + (\gamma - 1) h_i} \right)^{1/2}, \quad (2.4)$$

La répartition de l'échantillon parmi les strates qui minimise la variance (X_1) pour un coût global fixe prévu C^* (ou qui minimise le coût $E[C(D_1)]$ pour une variance fixe V^*) peut être précisée à une constante de proportionnalité près par la formule suivante

2.5 Répartition optimale pour X_1

$$E[C(D_1)] = c_0 \sum_{i=1}^H m_i (1 + (\gamma - 1) h_i). \quad (2.3)$$

et

$$E[C(D_0)] = mc_0 (1 + (\gamma - 1) h). \quad (2.2)$$

La somme $c_1 + c_2$ représente le coût d'une sélection d'échantillon "fructueuse" et c_0 représente le coût d'une sélection "infructueuse", alors, selon Wakseberg (1978), $\gamma = (c_1 + c_2)/c_0$ représente le ratio (quotient) du coût d'une sélection fructueuse sur une sélection infructueuse. Le coût total de la sélection de l'échantillon et de la détermination des valeurs de Y est une variable aléatoire dans les deux plans, D_0 et D_1 . Soient $C(D_0)$ et $C(D_1)$ qui représentent le coût total de la tenue d'une enquête selon les deux plans respectifs, on peut simplement démontrer que

Le taux de succès est le plus faible. Les plans tronqués dont il a été précédemment discuté peuvent être compris dans ce type général de plan à condition d'admettre la répartition d'aucune unité d'échantillon à la strate résiduelle et de se servir de l'écart quadratique moyen au lieu de la variance.

2.2 Notation de base

Supposons que la base BCR des numéros de téléphone a été cloisonnée en H strates basées sur un attribut d'une banque de 100 qui peut être déterminé à partir soit de la base des numéros de téléphone BCR ou de celle sur annuaire. Le choix des banques de 100 est quelque peu arbitraire; on pourrait envisager des banques de 10 à 500 numéros consécutifs. Pour la i^{e} strate, alors

P_i = la proportion de la base incluse dans la strate,
 h_i = la proportion des numéros de téléphone de la strate qui sont des NR (c.-à-d., le taux de succès),
 w_i = la proportion moyenne des NR dans les banques de 100 non vides (c.-à-d., le taux de succès moyen pour les banques non vides),
 z_i = la proportion de la population cible incluse dans la strate, et
 t_i = la proportion des banques de 100 de la strate qui ne contiennent aucun NR.

Le taux de succès moyen pour la base s'obtient par la formule $h = \sum_{i=1}^H h_i P_i$ et la proportion des banques de 100 vides de la base s'obtient par la formule $\bar{t} = \sum_{i=1}^H t_i P_i$.

Tableau 1

Les valeurs approximatives des paramètres de la base pour un plan à deux strates fondées sur la base BCR et la liste de l'annuaire Donnelly. La strate n° 1 se compose de tous les numéros de téléphone des banques de 100 dont au moins un numéro de téléphone est inclus dans la base sur la liste Donnelly; la strate n° 2 contient tous les autres numéros

Strate	Proportion de la population (P_i)	Proportion de succès (h_i)	Taux de succès des banques de 100 vides (t_i)	Taux de succès des banques non vides (w_i)
1	.3804	.9402	.5210	.0300
2	.6196	.0598	.0204	.9584
				.5371
				.4900

En général, on ne connaît pas avec certitude que les P_i . Les données d'un projet de recherches mixte du Bureau of Labor Statistics et de l'Université du Michigan ont servi à établir les valeurs approximatives des paramètres h_i et w_i pour les deux strates de l'exemple. Les valeurs des autres paramètres ont été calculées à partir de la relation algébrique $t_i = 1 - (h_i/w_i)$ et $z_i = h_i P_i/h$. Les approximations de tous les paramètres de la base pour le plan à deux strates sont indiquées au Tableau 1 ci-dessus.

méthodes de sélection de l'échantillon selon la strate visée. La section n° 3 traite de plusieurs de ces méthodes. La section n° 4 analyse les répercussions, sur l'efficacité du plan d'enquête, d'une répartition "non optimale" de l'échantillon. Le présent document se termine par une discussion comparative générale de la technique de Mitofsky-Waksberg et des plans stratifiés.

2. LE PROBLÈME DE LA RÉPARTITION DANS LES PLANS D'ENQUÊTE TÉLÉPHONIQUE STRATIFIÉE

2.1 Antécédents

Il est ici présupposé que la base fondamentale d'échantillonnage se compose de tous les numéros de téléphone produits en apposant des suffixes de quatre chiffres à la liste BCR des codes (combinaisons) indicatif régional - préfixe. En outre, il est présupposé que chaque ménage de la population cible est "lié" à un numéro de téléphone et un seul dans la base fondamentale d'échantillonnage (afin d'éviter les complications d'une probabilité inégale de sélection).

Nous posons également l'hypothèse d'un accès (possiblement seulement indirect) à une liste lisible par machine de numéros de téléphone tirés d'un annuaire. Il convient de noter que beaucoup de ménages ont choisi de tenir leur numéro de téléphone confidentiel et qu'une telle base de sondage tirée d'un annuaire ne contiendra pas tous les NR. De par leur nature, les listes fondées sur des annuaires ne sont pas à jour et omettent certains numéros présentement actifs tout en incluant d'autres qui ne sont plus des NR. Du point de vue des plans d'enquête, ces deux bases ont tendance à différer radicalement. La base BCR inclut tous les NR et procure donc une "couverture" complète de tous les ménages de la population cible, mais 20% environ seulement des numéros de téléphone y sont des NR. Ainsi, le "taux de succès" (et par conséquent l'efficacité d'échantillonnage) sera bien faible dans le cas d'un plan à composition aléatoire simple à partir de la base BCR. Par opposition, une base typique tirée d'un annuaire/d'une liste ne couvre qu'environ 70 pour cent des ménages de la population cible, mais le taux de succès atteint de 85 à 90 pour cent. En général, l'efficacité de l'échantillonnage dans un plan à composition aléatoire simple à partir d'une base tirée d'un annuaire/d'une liste est de beaucoup supérieure à l'efficacité possible à partir de la base BCR en appliquant la technique de Mitofsky-Waksberg. Malheureusement, dans bien des cas, le recours aux bases de sondage constituées à partir d'annuaires est exclu en raison du faible taux de couverture.

La notion fondamentale qui sous-tend la méthode de stratification proposée tient dans l'utilisation de l'information tirée de la base sur annuaire/sur liste pour répartir la base BCR en deux strates ou plus dont les taux de succès diffèrent, puis dans une répartition de l'échantillon parmi les strates de façon à minimiser les coûts (la variance) étant donné une variance (un coût) spécifique. Dans les paragraphes suivants, on appelle strate résiduelle la strate dont

d'une base dualiste dans laquelle un échantillon tiré des numéros inscrits est combiné à un échantillon à composition aléatoire d'après une estimation postérieure à la stratification. Si la couverture de la population est de moindre importance, des listes des numéros inscrits peuvent servir à déterminer des banques de 100 qui comprennent au moins un numéro de résidence inscrit et l'échantillonnage peut être restreint à ces banques. Survey Sampling Inc. (1986), et auparavant Stock (1962) et Sudman (1973) à partir d'annuaires par numéros, ont choisi des échantillons de numéros de téléphone d'un tel type de banque de 100. Il est évident qu'une telle méthode ne procure pas une couverture complète des ménages dont le numéro de téléphone n'est pas inscrit, mais elle peut améliorer sensiblement l'efficacité de l'échantillonnage. De fait, ces méthodes à "base tronquée" produisent des taux de numéros résidents comparables ou supérieurs à ceux associés à la technique de Mitofsky-Waksberg, et éliminent la tâche fastidieuse de remplacement des numéros non résidentiels. Malheureusement, pour bien des organismes d'enquête, la troncature de la base entraîne des lacunes de couverture inacceptables. Le présent document a pour but d'examiner des plans stratifiés sur la base BCR au titre d'option de rechange à la troncature de la base et aux plans de Mitofsky-Waksberg. On pourrait, comme exemple de stratification de la base, diviser la base BCR en deux strates: une strate "de haute densité", constituée des numéros résidentiels des banques de 100 qui contiennent un ou plus d'un numéro inscrit, et une strate de "faible densité" comprenant tous les autres numéros de la base BCR. Le "point limite" définissant les strates de haute et de faible densités relève quelque peu de l'arbitraire; un critère de définition de deux numéros inscrits ou plus pourrait réduire la possibilité que des banques de 100 soient involontairement incluses par suite d'une erreur de composition d'un numéro de téléphone. Cette façon de stratifier n'exige pas l'accès direct à tous les numéros inscrits. Le dénombrement des numéros inscrits, ou tout autre indicateur de la présence de numéros de téléphone inscrits au sein d'une banque de 100 obtenu d'un annuaire par numéros (dans les régions métropolitaines où un tel service existe) ou d'une liste commerciale pour l'ensemble du pays suffirait. Des études préliminaires indiquent qu'environ 50% des numéros de la strate à haute densité sont des NR, alors que ce n'est le cas que d'environ 2% dans la strate de faible densité. On peut exploiter la différence de coût évidente de l'échantillonnage sur l'une ou sur l'autre des strates en effectuant une répartition différentielle de l'échantillon. Les numéros de téléphone de la strate à faible densité pourraient faire l'objet d'une stratification ultérieure au moyen d'une analyse soignée des caractéristiques des banques de 100, à la lumière d'autres données disponibles de la base BCR et/ou de la liste Donnelley, ce qui pourrait se traduire par une efficacité d'échantillonnage encore plus grande.

La section suivante porte sur la répartition appropriée de l'échantillon parmi les strates dans le cas d'échantillonnage aléatoire simple pour chaque strate. Une caractéristique importante de la méthode d'échantillonnage téléphonique stratifié réside dans la possibilité d'appliquer diverses

La technique de Mitofsky-Waksberg présente par ailleurs plusieurs inconvénients. Par exemple, ce ne sont pas toutes les banques de 100 qui contiennent les k numéros résidents voulus, et il est donc possible de tenter d'obtenir les numéros aléatoires de deuxième degré en se servant des 99 autres chiffres sans atteindre les k NR nécessaires. En outre, il peut être difficile de déterminer si chaque numéro produit est bien celui d'une résidence, particulièrement à la phase de premier degré. Par exemple, beaucoup de régions rurales ne sont pas dotées du matériel qui avise la personne qui appelle lorsqu'un numéro n'est pas attribué. Les appels à des numéros non attribués sont transférés à une machine "à sonner". Il est difficile, dans ces régions, de distinguer les numéros non attribués des numéros de résidences où personne n'est présente lors de la période de l'enquête. Cette difficulté devient plus évidente à la fin de la période de l'enquête en raison de la nécessité de remplacer les numéros non résidentiels. On dispose alors de moins de temps pour appeler les numéros produits pour remplacer des numéros non résidentiels du deuxième degré de l'échantillonnage. Un petit résidu de numéros non réglés s'accumule et la détermination définitive de leur nature devient impossible en raison des contraintes de durée de l'enquête. Certains ont proposé des méthodes de traitement de ces numéros non réglés (Burkheimer et Levinsohn 1988), mais elles contrastent dans bien des cas la simplicité globale de la méthode.

On peut atténuer l'importance d'un bon nombre des lacunes associées à la technique de Mitofsky-Waksberg en procédant à un tri préalable des numéros de téléphone et en recourant aux systèmes d'entrées assistées par ordinateur. Toutefois, on n'élimine ces difficultés qu'en entravant la simplicité fondamentale de la méthode et/ou les principes d'échantillonnage probabilistique qui la sous-tendent (voir par exemple Porthoff 1987 et Brick et Waksberg 1991). D'autre part, des listes de numéros de téléphone inscrits ont servi de base d'enquête. On peut obtenir ces listes de numéros inscrits, pour l'ensemble du pays, d'entreprises commerciales comme Donnelley Marketing Information Systems. La sélection simple de numéros de téléphone, à partir de telles listes, procure un taux très élevé de NR (typiquement au moins 85%), mais ne touche malheureusement pas les ménages dont le numéro de téléphone n'est pas inscrit. La comparaison entre les ménages dont le numéro de téléphone est inscrit et ceux dont il ne l'est pas (voir par exemple, Brunner et Brunner 1971) révèle qu'un biais sensible peut en résulter.

Les listes des numéros de téléphone inscrits peuvent servir de façon à assurer également la couverture des ménages dont le numéro de téléphone ne l'est pas. Groves et Lepkowski (1986) décrivent une méthode de traitement

Plans d'enquête téléphonique stratifiée

ROBERT J. CASADY et JAMES M. LEPKOWSKI¹

RÉSUMÉ

Aux États-Unis, l'échantillonnage des numéros de téléphone de ménages s'appuie abondamment sur les méthodes de composition (téléphonique) aléatoire à deux degrés conçues par Mitofsky et ensuite élaborées par Waksberg. Par rapport à ces dernières, les autres méthodes courantes sont lacunaires au plan de la couverture et à celui des coûts. On remédie à ces lacunes par des plans d'échantillonnage qui, à partir de l'information sur les numéros inscrits, améliorent le rapport coût-efficacité de la composition aléatoire. La base de sondage, composée des numéros de téléphone, est divisée en deux strates, dont l'une comprend les numéros publiés pour lesquels l'information existe au niveau de la banque de 100, et l'autre, les numéros pour lesquels cette information n'existe pas. L'efficacité de divers modèles d'échantillonnage joints au plan stratifié est comparée à la composition aléatoire simple (simplement aléatoire) et à la technique de Mitofsky-Waksberg. On constate des gains d'efficacité dans le cas de presque tous ces modèles. Les hypothèses simplificatrices relatives aux valeurs des paramètres de la population de chaque strate se révèlent n'avoir que peu de répercussions globales sur l'efficacité estimée.

MOTS CLÉS: Composition aléatoire; répartition optimale; couverture; efficacité relative.

1. PLANS D'ENQUÊTE TÉLÉPHONIQUE: LA SITUATION COURANTE

Le plan à composition aléatoire à deux degrés pour échantillonner les téléphones des ménages, initialement proposé par Mitofsky (1970) puis élaboré par Waksberg (1978), a fait l'objet de nombreuses applications dans le cadre des enquêtes téléphoniques. La technique de Mitofsky-Waksberg tire parti d'une caractéristique de la distribution aux États-Unis des numéros résidentiels actifs (désignés ci-après NR), à savoir que les NR ont tendance à se regrouper en grappes denses dans des banques de numéros de téléphone consécutifs. De nos jours, dans l'ensemble des États-Unis, vingt pour cent environ seulement des numéros de téléphone qu'il est possible de dériver d'une combinaison connue d'un indicatif régional et d'un préfixe de trois chiffres sont des NR. Toutefois, si l'on peut déceler une banque de 100 numéros de téléphone consécutifs dont au moins un des numéros est un NR, alors, en moyenne, plus de 50 pour cent des numéros de la banque le seront également. Toute technique capable d'identifier les banques de 100 qui renferment des NR réduira grandement le fardeau du tri nécessaire à la détermination des numéros de téléphone attribués à des ménages. La première étape de la technique à deux degrés de Mitofsky-Waksberg consiste à obtenir une liste des combinaisons indicatif régional et préfixe (de trois chiffres) propres au territoire cible de l'enquête (disponibles de BellCore Research pour l'ensemble du pays; voir Lepkowski 1988). Une base de sondage comprenant les numéros de téléphone, ci-après désignée la base de BellCore Research ou base BCR, est alors constituée en associant les 10,000 suffixes de quatre chiffres (c.-à-d., 0000 à 9999) aux combinaisons indicatif régional et préfixe. On regroupe les numéros de

téléphone de la base d'enquête en banques de 100 (numéros), en se servant de l'indicatif régional, du préfixe de trois chiffres et des deux premiers chiffres du suffixe pour définir chaque banque. Par exemple, la combinaison indicatif régional et préfixe 313/764 donne 100 différentes banques de 100: 313/764-00, 313/764-01, ..., 313/764-99. Ensuite, on choisit un échantillon de banques de 100 et produit un seul numéro téléphonique complet pour chaque banque choisie en ajoutant un nombre aléatoire de deux chiffres au groupe de chiffres qui définit cette banque. Au premier degré de l'échantillonnage, on compose chacun des numéros de téléphone ainsi produits afin de déterminer et d'enregistrer s'il s'agit d'un numéro résidentiel. Toutes les banques de 100 dont le numéro obtenu par tirage aléatoire n'est pas un numéro résidentiel (NR) sont mises de côté. Un échantillon de second degré de NR est choisi à partir de toutes les banques de 100 dont le numéro (initialement) obtenu par tirage aléatoire était (est) un NR. Normalement, un nombre égal de numéros, disons k , sont alors produits dans chaque banque afin d'amorcer le processus d'échantillonnage de second degré. Si l'un de ces numéros de second degré se trouve être un numéro non résidentiel, on le remplace par un autre choisi au hasard dans la même banque. Le processus se continue jusqu'à l'obtention de k NR dans chaque banque. Il en résulte un échantillon de deuxième degré basé sur la sélection de banques de 100, les probabilités étant proportionnelles au nombre de numéros résidentiels de chaque banque. Cette méthode-logie s'est révélée à l'usage une excellente technique de détermination des banques de 100 renfermant des NR. La technique offre des avantages évidents. La proportion des numéros résidentiels au sein des banques de 100 retenues pour l'échantillonnage de deuxième degré est beaucoup plus élevée qu'au sein de la base BCR en général,

¹ Robert J. Casady, Bureau of Labor Statistics, U.S. Department of Labor et James M. Lepkowski, Survey Research Center, University of Michigan.

Tableau 5

Pourcentage d'augmentation de la variance inconditionnelle, I , pour la répartition proportionnelle quand G est estimé par \hat{G} . $C^* = 1,000$, $c' = 1$ et $c_1 = c_2 = 16$

G	\hat{G}								
	1/100	1/36	1/16	1/4	1	4	16	36	100
1/100	0.0	6.0	19.8	69.3	174.9	389.8	817.3	817.3	817.3
1/36	6.2	0.0	4.4	33.1	103.8	251.4	547.7	547.7	547.7
1/16	21.9	3.9	0.0	12.1	57.1	156.2	357.9	357.9	357.9
1/4	71.7	30.7	12.6	0.0	11.8	51.2	138.5	138.5	138.5
1	128.9	67.2	37.3	7.3	0.0	7.5	35.9	35.9	35.9
4	179.1	101.6	63.3	22.3	5.7	0.0	5.4	5.4	5.4
16	210.2	123.4	80.3	33.5	12.9	2.3	0.0	0.0	0.0
36	220.4	130.7	86.0	37.4	15.7	4.0	0.0	0.0	0.0
100	225.9	134.6	89.1	39.5	17.2	4.9	0.0	0.0	0.0

Note: I est défini dans (4.3), $G = S_Y^2/S_B^2$ et la fonction de coït est donnée par (1.3).

où G est la valeur correcte de S_Y^2/S_B^2 et \hat{G} est utilisé uni-
quement pour déterminer n' et n .

La valeur optimale de $\text{Var}(\hat{Y})$ (c.-à-d. celle obtenue
avec G) dans (2.4) peut être exprimée sous la forme

suivante

$$V_p(\hat{Y})_G = \frac{S_Y^2}{1} \left(\frac{C^*}{c'} + c + 2\sqrt{c'c/G} \right)$$

$$(4.2) \quad -\frac{1}{N} \left(1 + \frac{1}{G} \right).$$

Si $(1/N)(1 + 1/G)$ est négligeable, l'augmentation en
pourcentage de la variance attribuable à l'estimation de
 G , $I = 100\{V_p(\hat{Y})_G - V_p(\hat{Y})_G\}/V_p(\hat{Y})_G$, est, d'après
(4.1) et (4.2),

$$I = \frac{(1 - G) + \sqrt{c'/c} \{ \sqrt{G} - 2\sqrt{G} + (G/\hat{G}) \}}{(1 + \sqrt{G/c'})^2} \times 100. \quad (4.3)$$

Notons que (4.3) dépend seulement de G , \hat{G} et c/c' .

Nous présentons au tableau 5 les valeurs de I pour
 $C^* = 1,000$, $c' = 1$, $c_1 = c_2 = 16$ et neuf valeurs de G
et de \hat{G} . Les conclusions qui suivent sont fondées sur les
résultats du tableau 2.10.1 de Tredex (1989), qui incluent
des valeurs additionnelles de G et de \hat{G} . Tant que G se situe
dans l'intervalle $[G/4, 4G]$, l'utilisation de G pour trouver
(n' , n) produit un accroissement maximal de 15% de la
variance, et généralement moins. Si G se situe dans l'inter-
valle $[G/2, 2G]$, l'accroissement de variance attribuable
à l'erreur de détermination du paramètre est d'environ 4%
ou moins. Avec l'augmentation de G , l'accroissement de
variance associée à de tels intervalles (p. ex. $[G/4, 4G]$)
s'atténue. Cela tient au fait que si G est grand, $n' = n$, et
 \hat{G} et G donnent tous deux la même répartition. Une mani-
festation de ce résultat peut être observée dans le groupe
de zéros figurant à l'angle inférieur droit du tableau 5.
Quand G est petit, c.-à-d. lorsque la stratification est
bonne, la répartition de l'échantillon est plus sensible à une

erreur de détermination de G que quand G est grand. Ces
résultats sont peu influencés par les valeurs attribuées à
 $c_1 = c_2$. En bref, pour la répartition proportionnelle, des
erreurs assez importantes de détermination du paramètre
du plan (G) entraînent des augmentations de variance
relativement faibles.

BIBLIOGRAPHIE

- BOOTH, G., et SEDRANSK, J. (1969). Planning some two-
factor comparative surveys. *Journal of the American Statistical*
Association, 64, 560-573.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3^{ème} Ed.).
New York: John Wiley.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G.
(1953). *Sample Survey Methods and Theory*, (Vol. 1). New
York: John Wiley.
- HUGHES, E., et RAO, J.N.K. (1979). Some problems of
optimal allocation in sample surveys involving inequality
constraints. *Communications in Statistics - Theory and*
Methods A, 8(15), 1551-1574.
- RAO, J.N.K. (1973a). On double sampling for stratification and
analytical surveys. *Biometrika*, 60, 125-133.
- RAO, J.N.K. (1973b). On double sampling for stratification and
analytical surveys. *Biometrika*, 60, 669.
- SEDRANSK, J. (1965). A double sampling scheme for analytical
surveys. *Journal of the American Statistical Association*, 60,
985-1004.
- SRINATH, K.P. (1971). Multiphase sampling in nonresponse
problems. *Journal of the American Statistical Association*,
66, 583-586.
- SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S.,
et ASOK, C. (1984). *Sampling Theory of Surveys with*
Applications, (3^{ème} Ed.). Ames, IA: Iowa State University
Press.
- TREDDER, R.P. (1989). Some problems in double sampling for
stratification. Thèse de doctorat non-publiée, University of
Washington.

Nous avons obtenu des résultats semblables dans le cas de l'augmentation en pourcentage de la variance *conditionnelle* selon la répartition de Rao, $I_c = 100(V^{(c)O}/V^{(c)R})$, où $V^{(c)R}$ et $V^{(c)O}$ sont obtenues par l'estimation de $E\{\sum_{i=1}^L w_i^2 S_i^2(1/n_i - 1/n_i')\}$ en utilisant, respectivement, la répartition de Rao et la répartition optimale. Les résultats, fondés sur 200 répétitions de Monte Carlo et présentés sous forme de tracés en boîtes dans Tredet (1989, figures 2.8.2 et C.1-C.3), peuvent être résumés de la façon suivante. Pour *toutes* les spécifications de paramètres, les médianes des distributions de I_c sont voisines de 0. La plupart des valeurs de I_c sont faibles: environ 95% des spécifications de paramètres ont des distributions de I_c avec troisième quartile inférieur à 10%. Toutefois, occasionnellement, on note des valeurs élevées de I_c : environ 15% des spécifications de paramètres ont une valeur maximale de I_c supérieure à 20%.

On peut expliquer ces résultats, en partie, en définissant la taille d'échantillon de deuxième phase *optimale* dans la strate i par $n_i = \xi_i(n') \cdot n_i'$, où la dépendance de n_i à l'égard de n' observée est mise en évidence par l'inclusion de $\xi_i(n')$ avec $0 < \xi_i(n') \leq 1$. On peut ensuite trouver la répartition optimale en choisissant les $\xi_i(n')$ de façon à minimiser (pour n' fixe)

$$(3.8) \quad \frac{1}{L} \sum_{i=1}^L \frac{\xi_i(n')}{w_i S_i^2},$$

sous la contrainte $\sum_{i=1}^L c_i n_i' \cdot \xi_i(n') = C^* - c' n'$ (voir 2.9). Par contre, dans la répartition de Rao, pour n' fixe, on choisit les v_i de façon à minimiser

$$(3.9) \quad \frac{1}{L} \sum_{i=1}^L \frac{v_i}{w_i S_i^2},$$

sous la contrainte $n' \sum_{i=1}^L c_i w_i v_i = C^* - c' n'$, c.-à-d. un coût prévu fixe.

La minimisation de (3.8) plutôt que de (3.9) donnera une variance conditionnelle, et donc une variance inconditionnelle, plus faibles. Toutefois, si n' est grand, la différence entre (3.8) et (3.9) sera tenue.

3.4 Recommandations

Une fois établies des estimations raisonnables des paramètres du plan, il faut d'abord comparer le rapport des coûts, c/c' , avec les bornes inférieures, LB_P et LB_R , dans (3.5) et (3.7) pour voir s'il est préférable de recourir à l'échantillonnage double plutôt qu'à l'échantillonnage aléatoire simple. Ces évaluations doivent être faites avec soin, car une utilisation non fondée de l'échantillonnage double peut entraîner une *réduction* de la précision. Si l'on dispose de bonnes estimations des paramètres du plan, il est préférable d'utiliser la répartition de Rao plutôt que la répartition proportionnelle. Dans une situation où il importe de s'en tenir à un budget fixe, nous recommandons d'utiliser une modification de la méthode de Rao:

$$V_P(\hat{Y})_G = \frac{S_W^2}{1} \left(\frac{c'}{c' + \sqrt{c'G}} + c + \sqrt{c'G} \right)$$

L'analyse qui précède suppose que les erreurs de détermination des paramètres du plan ont un impact minime sur les répartitions de l'échantillon. Dans la présente section, nous évaluons, d'après un cas simple, l'effet sur $\text{Var}(\hat{Y})$ d'une erreur de détermination d'un important paramètre du plan. Dans le cas de la répartition proportionnelle, le choix de n' et de n dépend seulement de $G = S_W^2/S_B^2$, c' et c (voir (2.3)). Si l'on estime G par \hat{G} et qu'on substitue les valeurs résultantes de n' et de n d'après (2.3) et (2.1), on a

4. SENSIBILITÉ DES RÉPARTITIONS À L'ESTIMATION DES PARAMÈTRES DU PLAN

Chacune de ces méthodes exige la connaissance de certains paramètres du plan. Pour la répartition de Rao, les v_i optimaux exigent que les W_i et les S_i^2 soient déterminés. On peut voir dans (2.9) que pour la répartition optimale, les n_i optimaux dépendent des S_i^2 , mais non des W_i . Toutefois, le choix optimal de n' exige que les W_i soient précisés. Srinath (1971) et Rao (1973a) ont proposé une autre méthode qui exige la connaissance des S_i^2 , mais non des W_i . De toute évidence, la répartition de Rao est celle qui exige le plus de paramètres du plan connus, tandis que la méthode de Srinath est celle qui en exige le moins. Puisque le choix de n' est, de façon générale, robuste à l'égard des erreurs de détermination des paramètres du plan (voir, par exemple, Sedransk 1965, section 4.2.3), la méthode de la répartition optimale pourrait bien fonctionner dans les conditions pour lesquelles la méthode de Srinath a été élaborée.

Une autre solution consiste à utiliser la méthode de Rao pour trouver les valeurs "optimales" de n' et des v_i , puis à appliquer un algorithme pour arrondir et modifier les n_i ($n_i = v_i n_i'$) de façon à s'assurer que le budget soit respecté pour chaque échantillon. Malheureusement, il est difficile d'élaborer la partie de l'algorithme nécessaire pour se prémunir contre les dépassements de coûts. Toutefois, pour éviter les valeurs élevées de l'erreur proportionnelle dans la variance *conditionnelle* (c.-à-d. I_c) qui sont parfois observées, il faut utiliser les valeurs *optimales* de n' et des n_i .

Utiliser la méthode de Rao pour trouver la valeur "optimale" de n' . Puis, les n_i étant donnés, utiliser la méthode de la répartition optimale (c.-à-d. minimiser (2.9)) pour trouver les n_i . Cette méthode garantit que le budget sera respecté pour chaque échantillon, préserve l'essentiel du (faible) gain de précision procuré par la répartition optimale, et est facile à appliquer.

3.3 Répartition optimale et répartition de Rao

Pour comparer la répartition optimale avec celle proposée par Rao, nous avons considéré un large éventail de valeurs des paramètres du plan c' , S^2 et $\{(c_1, S^2_1, W_1, \dots, L)\}$. Nous avons posé $C^* = 1,000$ et examiné les cas $L = 2$ et 3. Les valeurs des paramètres du plan pour $L = 2$ sont données au tableau 3. Notons que pour ces exemples, $G = S^2_W/S^2_B$ varie entre 0.01 et 10.00. Nous supposons tout au long de l'évaluation que N est suffisamment grand pour que S^2/N dans (1.2) soit négligeable.

Tableau 3

Valeurs des paramètres du plan pour le cas $L = 2$ strates

Paramètre	Valeurs
c'	0.125, 0.250, 0.500, 1.000
c_1	1, 4, 16
c_2	16
W_1	0.5, 0.6, 0.7, 0.8, 0.9
S^2	70.4, 128, 704
S^2_1	1, 4, 16, 64
S^2_2	64

Note: Les 720 combinaisons des paramètres ci-dessus ont toutes été utilisées. En outre, nous avons étudié toutes les combinaisons des valeurs de c' , S^2 , et S^2_1 ci-dessus avec

- (a) $c_1 = 16$; $c_2 = 1, 4, 16$ et $W_1 = 0.5, 0.6, 0.7, 0.8, 0.9$,
(b) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 1, 4, 16$; $c_2 = 16$, et
(c) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 16$; $c_2 = 1, 4, 16$.

Pour nous assurer que les deux répartitions soient comparables, nous avons suivi les étapes ci-dessous pour chaque spécification des paramètres du plan.

1. Fixer une valeur unique de n' . Nous avons utilisé les deux valeurs de n' déterminées comme les meilleures selon (a) la méthode de Rao et (b) la répartition optimale. 2. Pour chacune des K répétitions de Monte Carlo ($K = 200$ ou 500), nous obtenons $n'_i = (n'_1, \dots, n'_L)$ et ensuite $n = (n_1, \dots, n_L)$ en utilisant la répartition optimale et $v = (v_1, \dots, v_L)$ en utilisant la méthode de Rao. Dans ce dernier cas, nous utilisons l'algorithme qui effectue les ajustements appropriés lorsque le côté droit de (2.7) dépasse 1 pour une ou plusieurs strates.

Puisque ni le n résultant de la répartition optimale ni le n résultant de la méthode de Rao ($n'_i = v_i n'_i$) ne sont nécessairement des entiers, nous arrondissons les n'_i et nous les ajustons de telle façon que pour chaque échantillon, le budget soit satisfait (au degré d'approximation découlant du fait que des valeurs entières de n' et de n sont utilisées). Nous avons constaté qu'en l'absence de ces ajustements, nous obtenions des résultats anormaux quand la variance de \bar{Y} selon la répartition de Rao était inférieure à la variance correspondante selon la répartition optimale. Cela se produisait lorsque le coût total associé à la méthode de Rao était supérieur à celui de la répartition optimale.

3. Pour obtenir des estimations, $V^{(c)O}$ et $V^{(c)R}$, des variances conditionnelles, $E_{n'}\{\sum_{i=1}^L w_i^2 S^2_i (1/n'_i - 1/n'_i)\}$, de Rao, nous avons utilisé la moyenne de $\sum_{i=1}^L w_i^2 S^2_i$ correspondant à la répartition optimale et à la répartition de Rao, nous avons utilisé la moyenne de $\sum_{i=1}^L w_i^2 S^2_i$ sur les K répétitions. Les estimations de la variance inconditionnelle, $\text{Var}(\bar{Y})$, dans (1.2) sont désignées par $V^{(u)O}$ et $V^{(u)R}$, où $V^{(u)R} = V^{(c)R} + (S^2/n')$. Pour évaluer la précision de ces estimations, nous avons estimé les erreurs-types et les coefficients de variation de $V^{(u)R}$ et $V^{(c)R}$. Toutes les erreurs-types étaient inférieures à 0.0022. Les coefficients de variation pour $V^{(u)R}$ et $V^{(c)R}$ étaient inférieurs à 0.0074 et 0.023 respectivement. Ainsi, $V^{(u)}$ et $V^{(c)}$ fournissent des estimations précises des variances inconditionnelles et conditionnelles. Nous présentons, au tableau 4, des estimations de l'augmentation en pourcentage de la variance inconditionnelle moyenne selon la répartition de Rao, $I_n = 100(V^{(u)R} - V^{(u)O})/V^{(u)O}$, pour certains des paramètres du plan énumérés au tableau 3. Nous présentons les résultats seulement pour la valeur de n' déterminée comme étant optimale par la méthode de répartition optimale. Ces résultats correspondent de près à ceux observés pour les autres spécifications du tableau 3, à ceux correspondant au cas $L = 3$ et à ceux qui utilisent la valeur de n' déterminée comme étant optimale par la méthode de Rao. Il est clair, d'après le tableau 4, que les gains de précision sont faibles, allant de zéro à environ 4%.

Tableau 4

Pourcentage d'augmentation, I_n , de la variance inconditionnelle moyenne $V^{(u)}$ pour la répartition de Rao en comparaison de la répartition optimale, pour certains paramètres du plan avec $S^2 = 70.4$, $S^2_2 = 64$, $c_2 = 16$ et $c' = 1$

S^2_1	G	c_1		
		16	4	1

a. $(W_1, W_2) = (.9, .1)$

64	10.000	0.0	0.4	1.4
16	0.419	0.1	0.1	0.1
4	0.166	0.1	0.1	0.4
1	0.116	0.1	0.3	0.8

b. $(W_1, W_2) = (.7, .3)$

64	10.000	0.0	0.7	3.6
16	0.760	0.0	0.2	0.7
4	0.455	0.1	0.3	1.4
1	0.394	0.0	0.7	0.9

c. $(W_1, W_2) = (.5, .5)$

64	10.000	0.0	1.0	4.1
16	1.316	0.0	0.4	0.9
4	0.934	0.0	0.6	1.8
1	0.858	0.0	0.2	0.0

Note: $I_n = 100(V^{(u)R} - V^{(u)O})/V^{(u)O}$. Voir la note du tableau 1 pour les définitions des coûts et des variances.

3.2 Comparaisons avec l'échantillonnage aléatoire simple

Pour permettre la comparaison avec la répartition de Rao et la répartition proportionnelle, supposons un échantillon aléatoire simple de taille n^* avec un coût prévu $n^* \sum_{i=1}^L W_i c_i = n^* c$ (voir (1.3)). Ainsi, pour un coût prévu fixe $C^*, n^* = C^*/c$ et

$$\text{Var}(\bar{y}_{n^*}) = S^2 \left(\frac{C^*}{c} - \frac{1}{N} \right) \equiv V_S, \quad (3.3)$$

où \bar{y}_{n^*} est la moyenne d'échantillon. En utilisant (2.4) et (3.3), on obtient

$$V_S - V_P = \frac{1}{c} \left\{ (c - c') S_B^2 - 2 S_B S_W \sqrt{c'/c} \right\}. \quad (3.4)$$

On peut montrer que $V_S - V_P \geq 0$ si, et seulement si,

$$\frac{c'}{c} \geq \left(\sqrt{G} + \sqrt{1+G} \right)^2 = LB_P, \quad (3.5)$$

où $G = S_W^2/S_B^2$. En utilisant (2.8) et (3.3), on obtient

$$V_S - V_R = \frac{C^*}{c} \left\{ S^2 - \left(\bar{S}_Y + S_B \sqrt{c'/c} \right)^2 \right\}, \quad (3.6)$$

où, encore une fois, on suppose que $v_i^0 \leq 1$ pour tout i (voir (2.7)). On peut voir facilement que $V_S - V_R \geq 0$ si, et seulement si,

$$\frac{c'}{c} \geq \frac{S_B^2}{(S - \bar{S}_Y)^2} = LB_R. \quad (3.7)$$

En pratique, on estimera LB_P et LB_R dans (3.5) et (3.7) et on les comparera avec le rapport des coûts, c/c' , pour décider s'il est avantageux d'utiliser l'échantillonnage double, avec répartition proportionnelle ou répartition de Rao, plutôt que l'échantillonnage aléatoire simple. Au tableau 2, nous présentons les valeurs de LB_P et de LB_R pour chacun des exemples cités au tableau 1. Nous donnons également pour $c = 1, 5$ et 25 les valeurs de R , soit le pourcentage de réduction de variance résultant du recours à une méthode d'échantillonnage double plutôt qu'à l'échantillonnage aléatoire simple. Comme nous l'avons indiqué plus haut, cet ensemble d'exemples représente une vaste gamme de conditions se prêtant à l'échantillonnage stratifié. Pour une valeur donnée de c , l'intervalle des valeurs de R_P et de R_R indique le large éventail de gains (par rapport à l'échantillonnage aléatoire simple) qui peuvent être obtenus. Bien qu'on ait en général $LB_P \geq LB_R$, on constate que $LB_P \gg LB_R$ pour plusieurs exemples. Les résultats indiquent la possibilité de gains élevés résultant de l'échantillonnage double, notamment avec la répartition de Rao, lorsque le rapport c/c' est élevé. Inversement, si c/c' est relativement peu élevé, les gains sont modestes et, dans certains cas, l'échantillonnage aléatoire simple est même préférable. Il importe par conséquent que LB_P, LB_R et c/c' soient estimés avec soin.

Tableau 2
Pourcentage de diminution de la variance pour la répartition proportionnelle (R_P) et la répartition de Rao (R_R) en comparaison de l'échantillonnage aléatoire simple, pour certains exemples classiques.

Référence	L	LB_P	LB_R	$c = 1$		$c = 5$		$c = 1$		$c = 5$		R_R
				R_P		R_P		R_P		R_P		
				25	5	25	5	25	5	25	5	25

Cochran (1977), p. 93	2	3.8	2.6	-177.9	11.9	45.7	-136.0	28.3	57.4	78.6	56.7	69.4
Hansen et coll. (1953), p. 205	3	6.1	1.1	-102.8	-6.1	26.4	-4.1	59.9	59.9	78.6	56.7	69.4
Sukhatme et coll. (1984), p. 118	4	3.7	2.7	-132.8	12.8	46.6	-105.3	26.5	56.7	78.6	56.7	69.4
Hansen et coll. (1953), p. 210	4	17.4	0.7	-127.7	-21.3	3.6	23.0	58.9	69.4	78.6	56.7	69.4
Cochran (1977), p. 111	7	6.8	4.5	-197.8	-9.8	23.3	-164.5	3.9	33.8	78.6	56.7	69.4
Hansen et coll. (1953), p. 202	8	103.4	5.6	-38.2	-14.1	-4.0	-23.7	-0.9	8.5	78.6	56.7	69.4
Hansen et coll. (1953), p. 202	11	76.4	1.7	-44.0	-15.6	-3.9	-11.0	13.7	23.8	78.6	56.7	69.4
Hansen et coll. (1953), p. 235	11	4.5	2.2	-105.8	4.0	37.9	-62.0	32.0	59.7	78.6	56.7	69.4
Hansen et coll. (1953), p. 202	12	24.5	4.0	-71.6	-19.9	0.2	-42.8	3.9	21.8	78.6	56.7	69.4

Note: En utilisant (2.4), (2.8) et (3.3), $R_P = 100(V_S - V_P)/V_S$, $R_R = 100(V_S - V_R)/V_S$, et (LB_P, LB_R) tels que définis en (3.5) et (3.7). Dans ces exemples, $c' = 1$ et C^* , le budget total pour chacune des méthodes, est le même qu'au tableau 1.

dépend du n' observé. Évidemment, un effort additionnel (c.-à-d. un échantillonnage de Monte Carlo) est nécessaire pour trouver la répartition optimale. Contrairement à la répartition optimale, la méthode de Rao permet la sélection des fractions de sondage de deuxième phase *avant* l'observation des n'_i (voir (2.7)). Se reporter aux sections 3 et 4 pour une analyse additionnelle.

3. COMPARAISONS

3.1 Répartition proportionnelle et répartition de Rao

En supposant que $v_i^0 \leq 1, i = 1, \dots, L$, et en utilisant (2.4) et (2.8), on peut montrer que

$$V_P - V_R = \frac{1}{S_c} \left(S_w - \frac{\sqrt{c}}{S_c} \right) \times$$

$$\left\{ 2S_B \sqrt{c'} + c \left(S_w + \frac{\sqrt{c}}{S_c} \right) \right\}, \quad (3.1)$$

où $S_c = \sum_{i=1}^L W_i S_i \sqrt{c'_i}$. En se rappelant que $c = \sum_{i=1}^L W_i c_i$ et en utilisant l'inégalité de Cauchy-Schwarz, on obtient $S_w - S_c / \sqrt{c} \geq 0$. Ainsi, comme prévu, $V_P - V_R \geq 0$. Si l'on définit $\bar{S} = \sum_{i=1}^L W_i S_i$ et $\bar{S}_y = \sum_{i=1}^L W_i S_i \sqrt{y_i}$ où $y_i = c_i / \sum_{i=1}^L W_i c_i$, et si l'on utilise (3.1), on peut montrer que

Le premier terme de (3.2) est la réduction de variance si tous les coûts d'échantillonnage sont égaux, tandis que le deuxième terme de (3.2) est la réduction si toutes les variances des strates sont égales. Comme prévu, si $c_i = c$ et $S_i^2 = S, V_P = V_R$.
 Nous présentons au tableau 1 les valeurs de $R = 100 (V_P - V_R) / V_P$ pour un ensemble d'exemples classiques, avec $c_i = c$. Les différentes colonnes donnent les caractéristiques des populations concernées ($L, S^2, S_w^2, G = S_w^2 / S^2$ et C^* , ainsi que les valeurs de R correspondant à $c/c' = 1, 2, 5$ et 25. Cet ensemble d'exemples représente une vaste gamme de conditions se prêtant à l'échantillonnage stratifié. Pour une valeur donnée de c , l'intervalle des valeurs de R indique le large éventail de gains qui peuvent être obtenus. Il ressort clairement du tableau 1 que la répartition de Rao peut produire d'importantes réductions de variance, même si les coûts d'échantillonnage des strates à la deuxième phase sont égaux et dans des situations où la stratification n'est pas particulièrement efficace (comme l'indiquent les valeurs élevées de G dans trois exemples). À mesure que c augmente, R affiche un taux de progression à peu près constant (voir le tableau 1).

Tableau 1
 Pourcentage de diminution de la variance, R , pour la répartition de Rao en comparaison de la répartition proportionnelle, pour un ensemble d'exemples classiques

Référence	L	S^2	S_w^2	G	C^*	pour $c' = 1$ et $c =$			
						1	2	5	25
Cochran (1977), p. 93	2	52,448	17,646	0.51	30	15.1	16.6	18.6	21.6
Hansen et coll. (1953), p. 205	3	2,835,856	1,467,632	1.07	1,000	48.7	55.1	62.3	70.9
Sukhatme et coll. (1984), p. 118	4	72,238	23,509	0.48	100	11.8	13.5	15.7	18.9
Cochran (1977), p. 111	7	619	343	1.25	1,000	11.2	11.7	12.4	13.7
Hansen et coll. (1953), p. 202	8	47,393	45,595	25.36	1,000	10.5	11.0	11.5	12.0
Hansen et coll. (1953), p. 202	11	47,393	44,974	18.59	1,000	22.9	24.1	25.4	26.7
Hansen et coll. (1953), p. 235	11	2,039,184	820,722	0.67	1,000	21.3	24.8	29.1	35.1
Hansen et coll. (1953), p. 202	12	47,393	40,252	5.64	1,000	16.7	18.3	19.8	21.6

Note: $R = 100 (V_P - V_R) / V_P$ avec V_P et V_R définis comme en (2.1) et (2.5) et C^* est le budget total. La fonction de coût est définie en (1.3) et les variances (S^2, S_w^2, G) sont définies en (2.3).

Dans le présent article, nous comparons trois plans d'échantillonnage double, qui diffèrent selon la façon dont les tailles des échantillons, c.-à-d. n' et les n_i , sont choisies. Nous comparons aussi ces méthodes avec un échantillon aléatoire simple ayant le même coût total fixe.

Les différents plans sont présentés à la section 2, tandis que les comparaisons font l'objet de la section 3. La section 4 présente les résultats d'un examen de l'effet, sur le choix des tailles d'échantillons, d'une erreur de détermination d'un important paramètre du plan.

2. PRÉSENTATION DES MÉTHODES

2.1 Répartition proportionnelle

Dans la méthode de répartition proportionnelle, $n_i = nw_i$ où $n = \sum_{i=1}^L n_i$. En utilisant (1.2), on peut montrer que la variance de \bar{Y} en vertu d'une répartition proportionnelle, V_p , est donnée par

$$V_p = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \left(\frac{1}{n'} - \frac{1}{N} \right) \sum_{i=1}^L w_i S_i^2, \quad (2.1)$$

où $w_i = N_i/N$ est la proportion de la population représentée par les unités de la strate i . En substituant $n_i = nw_i$ dans (1.3), on obtient le coût total prévu

$$\bar{C}_p = c'n' + cn, \quad (2.2)$$

où $c = \sum_{i=1}^L w_i c_i$. Si l'on choisit n' et n pour que (2.1) soit minimum à un coût total prévu fixe $\bar{C}_p = C^*$, on obtient

$$n' = \frac{c'}{C^*} + \frac{\sqrt{c'cG}}{C^*}, \quad (2.3a)$$

$$n = \frac{C^*}{c} + \frac{\sqrt{c'c/G}}{C^*}, \quad (2.3b)$$

où $G = S_W^2/S_B^2$, $S_W^2 = \sum_{i=1}^L w_i S_i^2$ et $S_B^2 = S^2 - S_W^2$.

En utilisant (2.3), on obtient

$$V_p = \frac{1}{C^*} \left\{ \left(c' + \sqrt{c'cG} \right) S_B^2 \right\}$$

$$+ \left(c + \sqrt{c'c/G} \right) S_W^2 \left\{ - \frac{N}{S^2} \right\}. \quad (2.4)$$

2.2 Répartition de Rao

Rao (1973a,b) suggère de poser $n_i = v_i n_i'$ où les v_i ($0 < v_i \leq 1$) sont des constantes fixées avant

l'échantillonnage. En utilisant cette répartition dans (1.2), on peut montrer que la variance de \bar{Y} en vertu de la répartition de Rao, V_R , est donnée par

$$V_R = \left(\frac{1}{n'} - \frac{1}{N} \right) S^2 + \frac{1}{n'} \sum_{i=1}^L w_i S_i^2 \left(\frac{1}{v_i} - 1 \right). \quad (2.5)$$

Le coût prévu correspondant, \bar{C}_R , est

$$\bar{C}_R = c'n' + n' \sum_{i=1}^L c_i v_i w_i. \quad (2.6)$$

Les v_i qui minimisent (2.5) sous la contrainte $\bar{C}_R = C^*$, satisfont à la condition suivante

$$v_i^0 = \frac{S_i \sqrt{c'}}{S_B \sqrt{c_i}}, \quad (2.7)$$

pourvu que le côté droit de (2.7) ne dépasse pas 1, quel que soit i . Sinon, un algorithme est nécessaire pour déterminer les v_i optimaux (voir Rao 1973a,b). Puisque Rao minimise la variance *inconditionnelle*, les v_i optimaux ne dépendent pas des n_i' observés. Une fois que les v_i sont déterminés, n' est obtenu de (2.6). Si l'on suppose que $v_i^0 \leq 1$ pour chaque i ,

$$V_R = \frac{1}{C^*} \left(\sum_{i=1}^L w_i S_i \sqrt{c_i} + S_B \sqrt{c'} \right)^2 - \frac{N}{S^2}. \quad (2.8)$$

2.3 Répartition optimale

La répartition optimale des tailles d'échantillon peut être obtenue par la minimisation de (1.2) directement. Pour n' et n' fixes, choisissons les n_i de façon à minimiser

$$\sum_{i=1}^L w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n_i'} \right), \quad (2.9)$$

en supposant un coût restant fixe, $C^* - c'n' = \sum_{i=1}^L c_i n_i$ et $n_i \leq n_i'$. Un algorithme est nécessaire pour déterminer les n_i optimaux, les n_i' étant données; voir Hughes et Rao (1979) et Tredet (1989). On peut trouver la valeur optimale de n' en évaluant (1.2) pour une suite de valeurs "d'essai" de n' . Pour chacune de ces valeurs, la valeur prévue de (2.9) est estimée au moyen d'un échantillonnage de Monte Carlo de n' , pour n' donné, est simple. Il y a plusieurs différences entre la répartition optimale et la répartition de Rao. Dans la première, les coûts totaux ne dépasseront pas C^* , tandis que dans la seconde, la répartition garantie seulement que le budget sera respecté en moyenne. Dans la seconde, les v_i sont fixes dans un échantillonnage répété, tandis que dans la première, la répartition des n_i

Echantillonnage double en vue d'une stratification

R.P. TREDER et J. SEDRANSKI¹

RÉSUMÉ

L'échantillonnage double est fréquemment utilisé en remplacement de l'échantillonnage aléatoire simple lorsque l'échantillonnage stratifié apparaît comme avantageux, mais que les unités ne peuvent être attribuées à des strates avant l'échantillonnage. Il est supposé, tout au long de l'article, que l'enquête a pour but d'estimer la moyenne d'une population finie. Nous comparons l'échantillonnage aléatoire simple et trois méthodes de répartition pour l'échantillonnage double: (a) répartition proportionnelle, (b) répartition de Rao (Rao 1973a,b) et (c) répartition optimale. Nous examinons également l'effet, sur le choix des tailles d'échantillon, d'une erreur de détermination d'un important paramètre du plan.

MOTS CLÉS: Tailles d'échantillon optimales; échantillonnage à deux phases.

1. INTRODUCTION

Supposons que nous voulions estimer la moyenne d'une population finie stratifiée, mais que les unités ne puissent être attribuées aux strates avant l'échantillonnage. Le plus souvent, le nombre d'unités de chaque strate est alors inconnu. Il est courant, dans un tel cas, de recourir à l'échantillonnage double comme solution de rechange à l'échantillonnage aléatoire simple. Avec l'échantillonnage double, un échantillon aléatoire simple de taille n' est prélevé dans une population finie de N unités, échantillon dont n'_i unités sont identifiées comme membres de la strate i , $i = 1, \dots, L$. L'échantillon de deuxième phase est constitué d'un ensemble de L sous-échantillons aléatoires simples indépendants résultant du prélèvement, dans la strate i , de n'_i unités parmi les n'_i identifiées à la première phase. Si y_{ij} désigne la valeur de Y pour la j -ième unité de l'échantillon de deuxième phase dans la strate i , la moyenne pour une population finie, \bar{Y} , est estimée par

$$\bar{\hat{Y}} = \sum_{i=1}^L w_i \bar{y}_i,$$

où $w_i = n'_i/n'$ et $\bar{y}_i = \sum_{j=1}^{n'_i} y_{ij}/n'_i$.

Soient $\sigma(n'_i)$ et $\sigma(n'_j)$ les ensembles de valeurs pour les unités des échantillons de première et de deuxième phase, respectivement, dans la strate i . Posons également $n' = (n'_1, \dots, n'_L)$ l'ensemble des valeurs pour toutes les unités de l'échantillon de première phase. Par ailleurs, soit $y_{n'}$ la moyenne des valeurs dans $\sigma(n')$, y'_i la moyenne d'échantillon de $\sigma(n'_i)$, $s_i'^2 = \sum_{j=1}^{n'_i} (y_{ij} - y'_i)^2 / (n'_i - 1)$ la variance d'échantillon de $\sigma(n'_i)$, $S_i^2 = \sum_{j=1}^{N_i} (X_{ij} - \bar{Y})^2 / (N_i - 1)$ la variance de population de la strate i et S^2 la variance de population finie analogue. Nous supposons toujours que n' est suffisamment grand pour que $Pr(n'_i = 0)$ soit négligeable. Puisque $1 \leq n_i \leq n'_i$, on a

$$C = c'n' + \sum_{i=1}^L c_i n_i, \quad (1.3)$$

Supposons la fonction de coût linéaire

$$+ E_{n'} \left\{ \sum_{i=1}^L w_i^2 S_i^2 \left(\frac{1}{n'_i} - \frac{1}{N} \right) \right\}. \quad (1.2)$$

$$= S^2 \left(\frac{1}{n'} - \frac{1}{N} \right)$$

$$+ E_{\sigma(n')} \left\{ \sum_{i=1}^L w_i^2 s_i'^2 \left(\frac{1}{n'_i} - \frac{1}{N} \right) \right\} \quad (1.1)$$

$$= V_{\sigma(n')}(\bar{Y}^{n'})$$

$$V(\bar{\hat{Y}}) = V_{\sigma(n')} E\{\bar{\hat{Y}} | \sigma(n')\} + E_{\sigma(n')} \{V(\bar{\hat{Y}} | \sigma(n'))\}$$

et

$$E(\bar{\hat{Y}}) = E_{\sigma(n')} \{E(\bar{\hat{Y}} | \sigma(n'))\} = \bar{Y}$$

où c' est le coût d'échantillonnage d'une unité à la première phase et c_i est le coût de la mesure de Y pour une unité de la strate i . Les tailles des échantillons, c.-à-d. n' et les n_i , sont choisies en fonction d'un coût total fixe ou d'un coût total prévu fixe.

¹ R. P. TREDER, Statistical Sciences, Inc. Seattle, Washington; J. SEDRANSKI, State University of New York at Albany, Albany, New York.

DREW, J.D., et GRAY, G.B. (1991). Standards and guidelines for definition and reporting of nonresponse to surveys. Prepared for the Second International Workshop on Household Survey Non-response, Washington, DC.

GOWER, A.R. (1979). Nonresponse in the Canadian Labour Force Survey. *Techniques d'enquête*, 5, 29-58.

GOWER, A., et ZYLSTRA, P.D. (1990). The use of qualitative methods in the design of a business survey questionnaire. Presented at the *International Conference on Measurement Errors in Surveys*, Tucson, Arizona.

HIDIROGLOU, M.A., et BERTHELOT, J.-M. (1986). Contrôle statistique et imputation dans les enquêtes-entreprises périodiques. *Techniques d'enquête*, 12, 79-89.

JULIEN, C., et MARANDA F. (1990). Le plan de sondage de l'enquête nationale sur les fermes de 1988. *Techniques d'enquête*, 16, 127-139.

KOVAR, J.G., MACMILLAN, J.H., et WHITRIDGE P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Direction de la méthodologie, document de travail, BSMO, 88-007E. Statistique Canada.

KUMAR, S., et DURNING, A. (1992). The Impact of Incentives on the Response Rates for FAMEX 1990: an Evaluation. Direction de la méthodologie, document de travail, SSMO, 92-001E. Statistique Canada.

LATOUCHE, M., et BERTHELOT, J.-M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8, 389-400.

LEMAITRE, G. (1983). Results from the Labour Force Survey Time and Cost Study. Rapport interne, Division des méthodes d'enquêtes ménage, Statistique Canada.

PLATEK, R., et GRAY, G.B. (1986). Sur les définitions des taux de réponse. *Techniques d'enquête*, 12, 19-30.

SWAIN, L., DREW, J.D., LAFRANCE, B., et LANCE, K. (1992). La création d'un registre des adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991. *Techniques d'enquête*, 18, 139-155.

REMERCIEMENTS

Les auteurs tiennent à remercier madame B.N. Chinnappa, de Statistique Canada, ainsi que les arbitres pour leurs précieux commentaires. Ils remercient également le Comité des méthodes et des normes de Statistique Canada pour ses suggestions et son appui dans l'élaboration du modèle de non-réponse.

BIBLIOGRAPHIE

BILLOU, F., et BERTHELOT, J.-M. (1990). Analysis on Grouping of Variables and on the Detection of Questionable Units. Direction de la méthodologie, document de travail, BSMO, 90-005E. Statistique Canada.

BILLOU, F., et FONTAINE, C. (1988). Etude sur la mise à la poste échelonnée pour le recensement des manufacturiers. Rapport du Statistique Canada.

CIALDINI R.B. (1991). Deriving Psychological Concepts relevant to survey participation from the literatures on compliance, helping and persuasion. International Workshop on Household Survey Non-response, Sweden, October 1990.

COLLEDGE, M.J. (1989). Coverage and classification maintenance issues in economic surveys. Dans Panel Surveys, (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons, 80-107.

CATLIN, G., et INGRAM, S. (1988). The effects of CATI on cost and quality. *Telephone Survey Methodology*, (Eds. R. Groves et al.). New York: Wiley, 437-450.

COUTTS, M., JAMIESON, R., WILLIAMS, B., et BRASLINS, A. (1992). The building of an integrated collection operation in Statistics Canada's regional offices. *Proceedings of the 1992 Annual Research Conference*. US Bureau of the Census, 395-411.

DREW, J.D. (1991). Recherche et essais pour les méthodes d'enquêtes par téléphone à Statistique Canada. *Techniques d'enquête*, 17, 63-75.

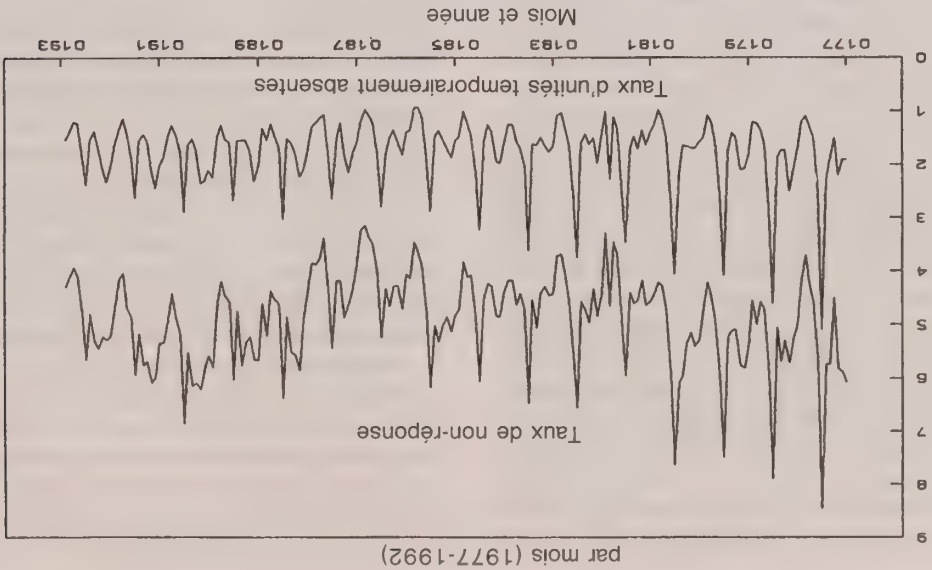


Figure 3. Taux de non-réponse à l'EPA

central au sein de l'organisme. Cette mesure facilitera l'analyse des tendances globales qui ont une incidence sur les taux de réponse et les taux de non-réponse des enquêtes. Nous avons discuté des mesures à prendre sous divers rapports du plan de sondage pour réduire au maximum la non-réponse et nous en avons illustré l'application pour deux grandes enquêtes périodiques. Bien que nous soyons limités au rôle qu'elles jouent à l'égard de la non-réponse, ces mesures constituent de bonnes pratiques d'enquête dont les avantages ne se limitent pas à de meilleurs taux de réponse.

Si l'on s'interroge sur l'avenir des taux de réponse dans les enquêtes effectuées au Canada, les tendances actuelles ne donnent pas lieu de s'inquiéter, malgré la légère augmentation des taux de non-réponse que nous avons constatée pour les enquêtes sociales dans la dernière décennie. Néanmoins, Statistique Canada poursuit ses efforts de recherche cognitive sur la non-réponse afin de mieux comprendre l'attitude et les préoccupations des répondants au sujet de questions comme la protection de la vie privée, la confidentialité, le fardeau de déclaration et le couplage d'enregistrements. En outre, des études sur la vérification sélective se concentrent actuellement sur les pratiques de vérification et de suivi pour de grandes unités. Il y a ainsi beaucoup de possibilités de réduire le fardeau de déclaration et les coûts, sans incidence marquée sur les estimations. Les conclusions de ces études nous permettront d'élaborer nos enquêtes et nos programmes statistiques de façon à respecter les préoccupations des répondants. Nous pourrions ainsi continuer de bénéficier de la grande collaboration de la population et des entreprises canadiennes.

L'étude du temps et des coûts effectuée en 1983 a eu lieu avant l'utilisation des interviews téléphoniques dans les petites zones urbaines et dans les zones rurales pour les unités faisant partie de l'échantillon depuis plus d'un mois, et avant le recours aux suivis téléphoniques pour les cas de non-réponse du premier mois. On envisage de répéter l'étude dans les conditions actuelles de l'enquête. Cette étude pourrait se pencher notamment sur la rentabilité des visites supplémentaires visant à réduire les taux de non-réponse. Bien qu'à partir de la quatrième, les visites supplémentaires représentent une proportion infime du nombre total de visites, leur part des coûts de la collecte de données peut être considérable en raison de la dispersion des logements en question. Les coûts de ces visites, ainsi que les renseignements sur les caractéristiques des logements visés par rapport aux caractéristiques des autres logements, permettraient d'évaluer le degré de suivi justifiable en tenant compte des coûts et de la variance.

5. RÉSUMÉ

Nous avons présenté ici des définitions normatives de la non-réponse. Dans une étude pilote portant sur sept grandes enquêtes-entreprises et enquêtes sociales de Statistique Canada, nous n'avons éprouvé aucune difficulté à appliquer ces définitions normatives. À compter de l'année de référence 1993, les données sur la non-réponse enregistrée selon ces normes dans les grandes enquêtes seront inscrites et conservées dans un registre

4.2 Enquête sur la population active

L'Enquête sur la population active (EPA) est la plus importante enquête permanente effectuée par Statistique Canada, la taille de l'échantillon étant d'environ 62,000 ménages par mois. Dans la section 3, nous avons étudié l'incidence de différents aspects du plan de sondage sur la non-réponse dans l'EPA. Dans la présente section, nous examinerons les tendances enregistrées antérieurement pour la non-réponse et nous approfondirons le rôle du suivi des non-répondants.

Le tableau 2 ci-dessous indique que pendant presque toute la période 1977-1991, les taux de non-réponse ont été stables, de l'ordre de 4 à 5%, de même que les taux de refus, de l'ordre de 1,0 à 1,5%. On peut néanmoins déceler quelques tendances. Par exemple, le Recensement de la population a une incidence favorable sur les taux de non-réponse enregistrés par l'EPA, ce qui démontre que la publicité qui entoure le Recensement profite aux enquêtes sur les ménages. Les taux de non-réponse ont chuté de 1 point de pourcentage entre 1980 et 1981, de 0,6 point entre 1985 et 1986 et de 0,4 point entre 1990 et 1991, seules années au cours desquelles on a enregistré une baisse importante des taux de non-réponse. En 1986, la diminution tenait presque entièrement à la baisse des taux de refus, tandis que cette baisse expliquait plus de la moitié de la réduction du taux de non-réponse en 1981. Si l'évolution des taux de non-réponse au cours de cette période est modérée, l'incidence positive du Recensement semble s'effriter graduellement. On observe, dans les quatre dernières années, une légère augmentation des taux de non-réponse et de refus, par rapport à la période allant de 1981 à 1987.

Le graphique ci-dessous (Figure 3) indique, par mois, les taux de non-réponse et d'unités temporairement absentes et illustre les tendances saisonnières des taux. Ainsi, les taux globaux de non-réponse atteignent un sommet pendant les mois d'été, parallèlement à une augmentation du taux d'unités temporairement absentes. Le graphique fait ressortir la forte corrélation entre le taux global de non-réponse et le taux d'unités temporairement absentes. La période de collecte de données de l'enquête s'étend habituellement sur six jours, soit la période de lundi à samedi qui suit la période de référence. Le samedi, les interviewers ont fait parvenir toutes données aux bureaux régionaux. Afin de réduire la pointe saisonnière, on a entrepris, vers la fin des années 1970, de faire un suivi le lundi pendant les mois de juillet et août. À l'occasion,

Tableau 2

Taux de non-réponse et de refus pour l'EPA, par année

	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
NR	5.42	5.39	5.35	5.37	4.41	4.67	4.65	4.57	4.69	4.08	4.23	5.07	5.18	5.57	5.20
REF	1.34	1.45	1.41	1.47	1.16	1.19	1.14	1.18	1.18	0.99	1.06	1.30	1.31	1.51	1.38

NR = Taux de non-réponse.
REF = Taux de refus.

En 1981, pendant le remaniement de l'EPA consécutif au Recensement, on a fait une étude détaillée du temps et des coûts. Cette étude avait pour objectif principal d'obtenir les renseignements nécessaires sur les coûts pour procéder à l'optimisation du rapport coût/variance de l'enquête. Elle a également permis de recueillir des renseignements intéressants sur les déplacements des interviewers et le déroulement des visites aux ménages (Lemaître 1983) et sur l'incidence du suivi des non-répondants sur les taux de réponse dans le cas des interviews directes. Lemaître a observé un taux de réponse de 92,4% après trois visites, soit le taux de réponse par visite étant régulièrement élevé, soit 56 à 61%, pour chaque série de visites. On a effectué un suivi plus poussé uniquement auprès de 3,5% des logements. Ces logements ont été visités encore 2,5 fois en moyenne et seulement 29% de ces visites ont permis d'obtenir une réponse. Les visites supplémentaires faites à ces ménages représentaient 5,8% de toutes les visites et ont porté le taux de réponse à 95,1%, soit une augmentation de 3,1 point.

Autre caractéristique notable de la tendance des non-réponses enregistrées dans l'EPA: le taux de non-réponse est plus élevé chez les ménages qui font partie de l'échantillon pour la première fois que chez les autres ménages. En 1980, le taux de non-réponse des ménages interviewés le premier mois était de 6,9%, contre 3,5% les mois suivants. Cet écart se manifeste surtout à l'égard des unités non contactées. Comme les interviewers effectuent surtout des interviews directes pendant le premier mois, leur nombre de tentatives de contact est limité. Les mois suivants, les interviews téléphoniques et les renseignements obtenus au cours de la première interview sur le moment le plus propice au contact permettent d'améliorer considérablement le taux d'unités contactées.

À partir de 1984, un changement s'est manifesté dans les pointes saisonnières du taux d'unités temporairement absentes. La pointe des mois d'été est moins forte, mais une seconde pointe en février et en mars devient plus prononcée. Cela semble indiquer que les ménages prennent de plus en plus souvent des vacances pendant l'hiver. Depuis quelques années, le suivi du lundi s'effectue donc également en mars si la semaine d'enquête coïncide avec la relâche scolaire.

le suivi du lundi s'étend à l'enquête de juin, selon la date de fin de l'année scolaire. Le suivi du lundi auprès des enquêtes qu'on n'a pu réjoindre pendant la semaine d'enquête s'effectue depuis les bureaux régionaux. Il en a résulté une baisse du nombre de cas de non-réponse attribuables à l'absence temporaire d'unités.

À partir de 1984, un changement s'est manifesté dans les pointes saisonnières du taux d'unités temporairement absentes. La pointe des mois d'été est moins forte, mais une seconde pointe en février et en mars devient plus prononcée. Cela semble indiquer que les ménages prennent de plus en plus souvent des vacances pendant l'hiver. Depuis quelques années, le suivi du lundi s'effectue donc également en mars si la semaine d'enquête coïncide avec la relâche scolaire.

Remaniée en janvier 1990, la nouvelle enquête diffère de l'ancienne – qui date du début des années 1970 – sous plusieurs aspects. Premièrement, afin d'accroître l'efficacité du plan de sondage, on a fait passer le nombre de groupes d'activité économique de 34 à 18 et on a utilisé trois strates de taille au lieu de deux. Deuxièmement, on a assoupli les niveaux de fiabilité. Ces modifications ont permis de réduire l'échantillon de 35%, ce qui a favorisé un suivi intensif des non-répondants. Troisièmement, la collecte de données a été confiée aux bureaux régionaux. Ce remaniement a fait monter les coûts unitaires de la collecte de données en raison du suivi supplémentaire; toutefois, la qualité des résultats s'est améliorée dans l'ensemble grâce à la réduction de la non-réponse.

La figure 2 présente, pour la période 1986-1992, les taux de réponse pondérés – provisoires et révisés, qui sont définis comme le rapport entre l'estimation des ventes attribuées aux répondants et l'estimation des ventes de toutes les unités incluses dans le champ de l'enquête. Ce graphique fait ressortir clairement que les taux de réponse provisoires et révisés de la nouvelle enquête sont considérablement plus élevés que ceux de l'ancienne. Les taux provisoires sont passés de 75% à 93%, tandis que les taux définitifs sont passés de 85% à 95%. En outre, l'écart entre les taux de réponse provisoires et révisés s'est nettement rétréci dans le cas de la nouvelle enquête. Il convient de souligner que les taux provisoires pour septembre 1991 étaient plus bas que prévu en raison d'une grève des employés de bureau chargés du traitement des documents. Plusieurs facteurs ont contribué à l'amélioration des taux de réponse, plus particulièrement le mode de collecte de données et les méthodes de suivi. Auparavant, le bureau

central (Division de l'industrie) se chargeait d'envoyer les questionnaires et ceux-ci lui étaient retournés une fois remplis. L'envoi postal s'effectuait selon des modalités de déclaration contrôlées manuellement. Le suivi immédiat auprès des non-répondants se limitait aux grandes unités et s'effectuait par téléphone à partir d'Ottawa. Le suivi des petites unités non répondantes se faisait un mois plus tard par courrier et se poursuivait, le cas échéant, pendant deux autres mois. Les unités qui n'avaient pas répondu au bout de trois mois étaient signalées aux bureaux régionaux en vue d'un suivi téléphonique.

Dans le cas de la nouvelle enquête, les unités de l'échantillon qui y participent pour la première fois (nouveaux participants) reçoivent une lettre préliminaire expliquant l'objet de l'enquête et l'importance de leur participation. Cette lettre est accompagnée d'un questionnaire vierge. En outre, environ une semaine après la date prévue de réception de cet envoi, on téléphone à chaque nouveau participant pour lui fournir des précisions, répondre à ses questions et lui offrir le choix entre une collecte de données par la poste ou par téléphone. Aux répondants qui choisissent la poste, la Division de l'industrie envoie les questionnaires selon des modalités de collecte automatisée qui sont établies à l'aide de l'information contenue dans le Registre des entreprises (RE). Ces modalités de collecte sont mises à jour dans le RE à l'aide de profils établis par la Division du registre des entreprises et de nouveaux renseignements obtenus par les bureaux régionaux auprès des répondants. Si les enquêtes préfèrent répondre par téléphone, les bureaux régionaux les interviewent aux dates et aux heures convenues au préalable et transmettent les données recueillies au bureau central à la fin de chaque cycle de collecte mensuel.

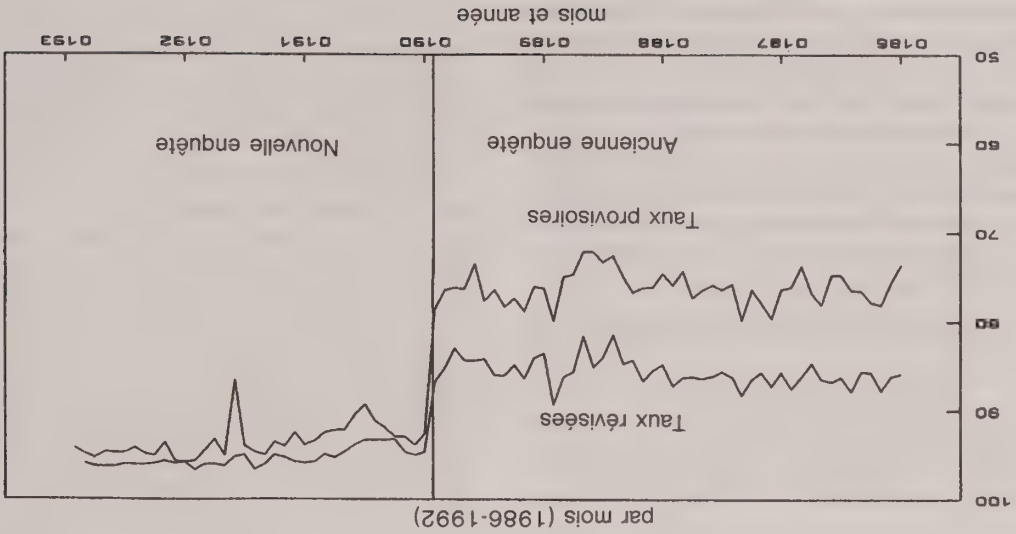


Figure 2. Taux de réponse à l'EMCD

A l'égard des enquêtes-entreprises et des enquêtes sur l'agriculture de Statistique Canada, on allège le fardeau

3.5 Prise en compte de données administratives

des enquêtes sur l'agriculture. On repère ainsi les enregistrements qui ont une incidence appréciable sur les estimations, et on limite le suivi à ces enregistrements. Les enregistrements qui ont une faible incidence sont soumis à un processus automatisé de contrôle et d'imputation en vue d'assurer la cohérence des résultats. Troisièmement, pour réduire au maximum le fardeau de déclaration, il faut repérer toutes les erreurs relatives à chaque unité qui doit faire l'objet d'un suivi afin de corriger la majorité des erreurs au cours d'une seule interview. À cette fin, un module d'analyse des contrôles inter-zones et de localisation des erreurs comme celui contenu dans le Système généralisé de vérification et d'imputation mis au point par Statistique Canada, peut s'avérer très utile (Kovar, MacMillan et Whitridge 1988). Lorsque trop de données sont rejetées à la vérification mais s'avèrent exactes à l'étape du suivi, il convient de modifier les règles de contrôle pour alléger le fardeau de déclaration.

Les méthodes de vérification sélective des données numériques élaborées par Statistique Canada peuvent être regroupées en trois ensembles: i) vérification statistique, ii) groupement de variables et iii) fonction de caractérisation. En matière de vérification statistique, Hidiroglou et Berthelot (1986) ont élaboré un processus de transformation qui permet surtout de détecter les unités qui présentent des changements inhabituels d'une fois à l'autre. Ce processus tient compte du fait que, d'une période à l'autre, les petites unités subissent fondamentalement des changements plus variables que les grandes unités. Les limites des rejets à la vérification prennent donc la forme d'un entonnoir, ce qui rend possible des changements relatifs importants dans les petites unités. Ces limites se calculent à l'aide de médianes et de quartiles et sont donc robustes à l'égard des observations aberrantes. Cette méthode peut aussi servir à déceler des rapports de valeurs aberrantes entre deux variables. Toutefois, le nombre de comparaisons par paire peut devenir excessif. Bilocq et Berthelot (1990) ont proposé une méthode qui consiste à grouper les variables analogues en sous-ensembles, puis à vérifier les variables par recouplements à l'intérieur des sous-ensembles. Ce procédé de cloisonnement repose sur les méthodes de corrélation des composants principales. L'importance des erreurs, mesurée d'après leur incidence sur les estimations, entre également en ligne de compte. Dans le cas d'un rejet à la vérification de questionnaires remplis, Latouche et Berthelot (1992) ont élaboré une fonction de caractérisation qui attribue à chaque répondant un indice relatif de l'importance de l'erreur selon la taille de l'unité, le nombre de réponses douteuses dans le questionnaire et leur poids ainsi que l'importance relative des variables. Une étude simulant cette fonction de caractérisation a montré que le fait de recombiner avec un nombre restreint d'unités suffisait pour assurer une qualité acceptable pour les estimations finales.

L'Enquête mensuelle sur le commerce de détail (EMCD) consiste à recueillir des données sur les ventes d'un échantillon d'établissements de détail et sur les stocks d'un sous-ensemble de ces établissements. On établit des estimations de niveau et de variation pour ces deux variables. Le plan d'échantillonnage consiste dans un échantillon aléatoire simple avec renouvellement, stratifié selon la province, le secteur d'activité et le revenu commercial brut. La population compte environ 165,000 entreprises et l'échantillon, environ 13,000. Les données sont recueillies par téléphone auprès d'environ 40% des unités et par courrier auprès des 60% restants. Des estimations provisoires sont publiées sept semaines après la période de référence de l'enquête et les estimations finales, qui visent un plus grand nombre de répondants grâce au suivi des non-réponses, sont publiées un mois plus tard.

4.1 Enquête mensuelle sur le commerce de détail

Nous examinerons brièvement les taux de non-réponse pour deux enquêtes de Statistique Canada afin d'illustrer quelques facteurs généraux qui influent sur la non-réponse et que nous avons décrits à la section 3.

4. ANALYSE DES TAUX DE NON-RÉPONSE POUR CERTAINES ENQUÊTES

On doit disposer d'un bon système de surveillance afin d'établir en tout temps l'état du processus de collecte de données. Dans le cas des enquêtes-entreprises, les codes d'état de la collecte, qui sont consignés pour chaque unité enquêtée, servent à contrôler le processus de collecte de données. Ces codes d'état de la collecte, enregistrés dans l'ordre chronologique de l'exécution de l'enquête, s'utilisent avec d'autres codes qui indiquent l'état des activités de l'unité (entreprise active, saisonnière [avec dates d'exploitation], inactive, fermée provisoirement, etc.). Voici quelques exemples de codes d'état de la collecte: i) mode de collecte de données à différentes étapes du processus, ii) codes d'entrée en contact pour les unités (qui sont réputées actives pendant la période de référence) et pour les exclusions (entreprises fermées, inactives ou provisoirement fermées) et iii) dates prévues de retour des renseignements en vue d'un suivi supplémentaire. Le système de gestion reçoit de sources extérieures à l'enquête des renseignements qui indiquent le changement de l'état des unités et retracer l'état de la collecte de données depuis la collecte initiale jusqu'au suivi et ce, jusqu'à ce que toutes les unités soient classées dans l'une ou l'autre des catégories prévues selon le cadre décrit à la section 2.

3.6 Système de gestion de la collecte de données

de déclaration en obtenant de sources administratives une partie des données sur les petites unités. Ces données servent également à remplacer des données illisibles, incohérentes ou manquantes. Par exemple, les données qui auraient dû être produites par les petites unités non répondantes sont imputées à l'aide de données fiscales.

menés en 1992 afin de connaître l'incidence de l'IAO sur les estimations et sur la qualité des données, y compris les taux de réponse.

3.3.7 Incitatifs

En vertu de la Loi sur la statistique, qui définit les pouvoirs juridiques de Statistique Canada, la participation aux enquêtes de Statistique Canada est obligatoire pour les entreprises et les particuliers choisis, sauf si l'enquête est désignée facultative par le statisticien en chef. Un exemple d'enquête à participation obligatoire est le recensement de la population, où un refus catégorique de répondre peut entraîner des poursuites. Dans le cas des autres enquêtes, l'organisme cherche à obtenir la collaboration des répondants éventuels en leur envoyant des documents préliminaires ou de la documentation publicitaire qui expliquent l'objet de l'enquête et la confidentialité des données; l'organisme compte aussi sur l'entre-gent de ses représentants, notamment les interviewers qui se rendent sur place et qui ont pour instruction de présenter leur insigne d'identité et d'informer les répondants de l'objet de l'enquête et de la confidentialité des données.

On a effectué, à l'égard des enquêtes sociales, plusieurs études sur l'utilisation d'incitatifs. La première visait l'Enquête sur la population active (Gower 1979). Lors d'un essai portant sur un échantillon fractionné, la moitié des ménages ont reçu la publication "Le Canada" au moment du premier contact. Au cours des mois suivants, le taux de refus a été légèrement inférieur chez les ménages de l'échantillon qui avaient reçu la publication. Selon les interviewers, l'incitatif représentait un avantage négligeable et les méthodes existantes de visite à domicile étaient plus efficaces pour réduire le taux de non-réponse. Plus récemment, lors d'une étude sur les incitatifs à l'égard de l'Enquête de 1990 sur les dépenses des familles, les interviewers ont adopté trois démarches: i) remettre à chaque ménage choisi une plaquette à pince portant le logo de Statistique Canada; ii) remettre la publication "Portrait du Canada" de Statistique Canada; iii) ne remettre aucun incitatif aux ménages d'un échantillon de contrôle. Dans l'ensemble du pays, aucun changement significatif n'a été observé dans les taux de réponse (Kumar et Durning 1992). On prévoit faire une autre étude sur les incitatifs à l'égard d'une prochaine enquête longitudinale sur le revenu et le travail.

3.4 Vérification sélective

Une mauvaise vérification des données constitue une autre cause possible de non-réponse. Dans ces circonstances, il arrive que l'on recommande plusieurs fois avec le répondant pour le même questionnaire, ce qui le rend moins disposé à collaborer à l'avenir. Afin de rationaliser et d'optimiser le processus de vérification de manière à réduire au maximum les contacts répétés, il convient d'adopter les trois mesures suivantes. Premièrement, il doit y avoir cohérence entre les étapes de la vérification à la saisie, du suivi et de l'imputation. Deuxièmement, on effectue une vérification sélective des données numériques et particulièrement dans le cas des enquêtes-entreprises et

et pour tenter d'obtenir une interview. Bien que le suivi soit nécessaire pour abaisser le taux de non-réponse, il n'est plus rentable au-delà d'un certain point. Peu d'études ont cherché à déterminer quelles seraient les méthodes les plus appropriées pour établir un calendrier des suivis en prenant en considération les coûts et l'erreur totale. L'examen de cette question nécessiterait des études de coût pour estimer les paramètres d'un modèle de coûts et de variance. Ces paramètres seraient les suivants: tentatives de contact, résultats, coûts, et caractéristiques des répondants à différentes étapes du suivi. L'automatisation croissante de la collecte de données au cours des années à venir devrait faciliter la collecte et l'utilisation de ces renseignements en vue d'optimiser les méthodes de collecte de données.

3.3.6 Technologie

Dans le cas des enquêtes-entreprises, la collecte de données s'effectue surtout selon la méthode "papier et crayon". Signaux deux exceptions: l'Enquête mensuelle sur les industries manufacturières, qui utilise actuellement l'ITAO (Coutts *et al.* 1992), et l'Enquête annuelle sur les industries manufacturières, où l'on a fait l'essai de l'ITAO pour recueillir des données auprès des petits fabricants. Etant donné le succès de l'opération "ITAO" dans le cas de l'Enquête mensuelle sur les industries manufacturières, il est prévu d'adopter l'ITAO pour d'autres enquêtes-entreprises. On s'apprête également à mettre à l'essai d'autres méthodes de collecte de données pour les enquêtes-entreprises. Il s'agit de l'ordinateur de poche pour l'Indice des prix à la consommation, du carnet grille pour l'Enquête trimestrielle sur le camionnage pour compte d'autrui, et de l'introduction de données au moyen d'un téléphone à clavier pour l'Enquête sur l'emploi, la rémunération et les heures de travail.

La collecte de données pour les enquêtes sociales se fait aussi selon la méthode "papier et crayon", mais on a décidé de passer à l'interview assistée par ordinateur (IAO) dans les prochaines années. Les interviewers sur le terrain disposeront d'ordinateurs portatifs pour faire des interviews directes ainsi que des interviews téléphoniques depuis leur domicile. Cette décision a été prise à la suite de deux essais d'IAO dans l'EPA qui ont donné des résultats positifs. Le premier essai (Cattin et Ingram 1988) a révélé une amélioration de la qualité des données, par exemple une meilleure énumération des occupants des logements échantillonnés, et une réduction du nombre de rejets à la vérification, sans aucune incidence repérable sur les estimations ou sur les taux de réponse. Le second essai, effectué en 1991 (Coutts *et al.* 1992), a démontré l'utilité des ordinateurs portatifs pour les IAO effectuées par les interviewers sur le terrain. L'introduction de l'IAO dans les enquêtes sociales est prévue dès 1993. Il faudra attendre toutefois les résultats d'essais plus poussés qui auront été

par interviews téléphoniques au cours des mois ultérieurs. Au moment du premier contact avec le ménage, l'interviewer présente son insigne d'identité, explique au répondant l'objet de l'enquête et l'assure de la confidentialité des réponses avant de procéder à l'interview. Cette visite est précédée d'une lettre dans laquelle le directeur régional annonce au ménage qu'il a été choisi pour l'enquête et décrit le but de l'enquête. Les répondants sont invités à composer un numéro sans frais pour obtenir des précisions même pendant la tenue de l'enquête.

Dans le cadre d'un programme de recherche et d'essai pour des méthodes d'enquête téléphonique réalisé de 1985 à 1989, on a étudié la possibilité de remplacer la première interview directe par une interview téléphonique. La conduite de l'EPA sous forme d'enquête téléphonique menée depuis le bureau central a entraîné une augmentation des taux de non-réponse de l'ordre de 68 à 75%. On a également constaté une augmentation du biais de non-réponse découlant de différences dans les caractéristiques de la population active chez les répondants et chez les non-répondants supplémentaires (Drew 1991). La seule enquête-ménages périodique de Statistique Canada à n'utiliser que les méthodes d'enquête téléphonique pour la collecte de données est l'Enquête sociale générale (ESG) annuelle, qui a recours à un sondage téléphonique au hasard portant sur 10,000 ménages. À l'occasion, on a ajouté à l'échantillon de l'ESG des ménages supprimés de l'EPA par renouvellement. Par exemple, un échantillon de personnes âgées déjà visé par l'EPA a été choisi pour les besoins de l'enquête sociale lorsque ce groupe d'âge présentait un intérêt particulier.

3.3.4 Conception du questionnaire et documents de présentation

La conception judicieuse du questionnaire favorise non seulement l'exactitude des données recueillies, mais aussi l'augmentation des taux de réponse. Le questionnaire et la documentation de présentation sont particulièrement importants dans le cas des enquêtes postales, puisqu'ils constituent le seul contact avec le répondant. La documentation envoyée au répondant doit expliquer l'objet de l'enquête, mentionner l'instance en vertu de laquelle cette enquête est effectuée, garantir la confidentialité des réponses et donner le numéro de téléphone d'une personne de l'organisme pouvant donner des précisions sur le questionnaire. Les questionnaires doivent être soumis à un processus de révision indépendant de leur conception. Ce processus consiste en des examens effectués par des spécialistes de l'organisme ou par des groupes de discussion. Le recours à des groupes de discussion ou à la recherche cognitive a donné lieu à plusieurs améliorations qui ont contribué à accroître la motivation des répondants et a eu aussi pour conséquence de simplifier la tâche de remplir les questionnaires de plusieurs enquêtes de Statistique Canada, notamment le Recensement de la population, l'Enquête sur la population active, le Recensement sur l'industrie de la construction et l'Enquête sur l'emploi, la rémunération et les heures de travail (Gower 1990).

3.3.5 Stratégies de suivi

Dans le cas des enquêtes-entreprises et des enquêtes sociales, le suivi fait partie intégrante du plan de sondage. Ce n'est que grâce à un suivi intensif qu'on peut maintenir à un faible niveau le taux de non-réponse attribuable aux non-contacts. Comme le suivi est habituellement plus coûteux par unité que la collecte initiale (en supposant que le coût de l'enquête est fixe), l'envergure du suivi a une incidence directe sur la taille de l'échantillon, donc sur la variance, et sur le taux de réponse, donc sur le biais de non-réponse. Les stratégies d'élaboration vont du vaste échantillon accompagné d'un suivi restreint au petit échantillon accompagné d'un suivi intensif. Lors du remaniement de l'Enquête sur le commerce de gros et de détail dans les années 1980, on cherchait en priorité à améliorer les taux de réponse; on a donc opté pour la stratégie du petit échantillon accompagné d'un suivi intensif.

En ce qui concerne les enquêtes-entreprises, le suivi sert à la fois à obtenir des données auprès des non-répondants et à communiquer de nouveau avec des répondants dont les données ont été rejetées à la vérification. La plupart des enquêtes-entreprises utilisent le courrier comme mode principal de collecte car c'est un moyen peu coûteux qui permet aux entreprises de répondre en consultant leurs dossiers. Afin de réduire les coûts, on limite souvent le suivi des non-réponses à un sous-échantillon de non-répondants. La répartition et la sélection des unités non-répondantes reposent habituellement sur les facteurs suivants: i) une strate à tirage complet d'unités nécessitant un suivi, en vue de concentrer les efforts sur les grandes unités non répondantes; ii) une égalisation des taux de réponse dans toutes les strates; et iii) le renouvellement des petites unités non répondantes devant faire l'objet d'un suivi. Dans le cas des enquêtes intra-annuelles, le suivi des non-réponses s'effectue en général par téléphone, les contraintes de temps ne permettant pas un suivi par courrier. Dans le cas des enquêtes annuelles, où la rapidité de la collecte est moins critique, on a tendance à utiliser le courrier à la fois pour la collecte initiale et pour les premières tentatives de suivi des non-réponses, puis le téléphone en dernier recours. Depuis quelques années, cependant, le suivi s'effectue de plus en plus souvent par téléphone dans le cas des enquêtes annuelles.

Pour ce qui est des enquêtes sociales, la distinction n'est pas aussi nette entre collecte initiale et suivi des non-réponses. Le suivi consiste plutôt dans des tentatives ultérieures pour communiquer avec les ménages et les interviewer pendant la période de l'enquête. Il existe certaines distinctions, selon la situation du logement. On visite une première fois les logements nouvellement échantillonnés pour repérer ceux qui sont hors du champ de l'enquête et pour tenter de faire une interview directe avec les occupants des logements compris dans le champ de l'enquête. Lorsqu'il est impossible d'obtenir une interview, l'interviewer tente d'obtenir auprès d'un voisin les renseignements suivants: nom, numéro de téléphone et moment le plus propice auquel appeler. Les interviewers ont pour le plus propice de tenter encore deux ou trois fois d'obtenir une interview, soit par téléphone, soit en personne. En règle

en s'inspirant des travaux de Cialdini (1991). L'objectif consistera à cerner les méthodes utilisées par les interviewers de calibre supérieur afin de les enseigner aux autres interviewers.

3.3.3 Mode de collecte de données

Statistique Canada fait tout en son pouvoir pour offrir aux répondants le choix du mode de déclaration qui leur convient le mieux, y compris le choix d'une des deux langues officielles. Une telle souplesse permet d'améliorer les taux de réponse.

On peut classer les enquêtes-entreprises effectuées par Statistique Canada en deux grands groupes: les enquêtes annuelles et intra-annuelles. Dans le cas des enquêtes annuelles, les données sont recueillies surtout par l'envoi, d'Ottawa, de questionnaires à retourner par la poste, certains répondants fournissant leurs données sur bandes magnétiques ou sur disquettes. Le moment choisi pour l'envoi des questionnaires d'enquêtes-entreprises annuelles doit concorder avec la date de clôture de l'exercice du répondant pour fins fiscales, parce que les données nécessaires sont déjà disponibles à cette date et que l'ambiguïté relative à l'année de référence est réduite au maximum. Dans une étude sur le Recensement annuel des manufactures, Billocq et Fontaine (1988) ont constaté qu'on obtenait les meilleurs taux de réponse en communiquant avec les répondants trois mois après la fin de leur exercice financier. Il faut donc échelonner les envois en tenant compte des dates de clôture des exercices financiers. Dans le cas des enquêtes-entreprises intra-annuelles, la collecte de données s'effectue surtout par l'envoi, depuis le bureau central, de questionnaires à retourner aux bureaux régionaux. Quant aux unités non contactées par courrier, la plupart répondent par téléphone aux bureaux régionaux, tandis qu'un petit nombre de répondants envoient directement à Ottawa des réponses sur support informatique. Il est important de respecter les pratiques comptables des répondants. La plupart des répondants utilisent le mois civil pour leur comptabilité, tandis que d'autres utilisent des cycles de quatre ou de cinq semaines. Dans les deux cas, l'organisme d'enquête reçoit habituellement les données une ou deux semaines après la fin de la période mensuelle. Les interviews téléphoniques des enquêtes-entreprises servent à recueillir des données pour diverses raisons: précisions à apporter aux instructions, mesures de suivi, etc. Mal employé, ce mode de collecte risque de compromettre la qualité des réponses. Par exemple, un répondant peut être obligé d'estimer les données s'il ne dispose pas de dossiers près du téléphone. Si les interviews téléphoniques s'effectuent périodiquement, comme dans le cas d'enquêtes mensuelles, le fait de convenir avec le répondant d'une journée et d'une heure propices améliore les taux de réponse ainsi que la qualité des réponses.

Dans le cas des enquêtes sociales, dont l'Enquête sur la population active, le mode de collecte consiste en des interviews téléphoniques "préparées": on procède d'abord à une interview directe pendant le premier mois au cours duquel le ménage fait partie de l'échantillon, puis surtout

données, d'où son incidence positive sur les taux de réponse. Avant l'apparition des enquêtes téléphoniques, cette organisation décentralisée constituait la seule façon de mener les enquêtes sociales. De 1985 à 1989, on a mis en oeuvre un programme de recherche et d'essai pour des méthodes d'enquête téléphonique (Drew 1991), dans lequel on cherchait à établir un mode d'organisation mixte. Selon ce mode d'organisation, le rôle des interviewers locaux consisterait essentiellement à effectuer des interviews directes, les interviews téléphoniques étant effectuées depuis les bureaux régionaux. Ce mode d'organisation offrirait moins de possibilités de suivi direct auprès des ménages que l'on ne pourrait rejoindre par téléphone, d'où un taux de non-réponse un peu plus élevé. De plus, ce mode d'organisation entraînerait des frais généraux plus élevés, à cause de l'addition de locaux et de matériel de bureau dans les bureaux régionaux. Cela aurait pour conséquence de réduire de beaucoup le personnel sur le terrain, ce qui limiterait la capacité d'effectuer de grandes enquêtes spéciales qui nécessitent des interviews directes. En outre, ce mode d'organisation réduirait le bassin de personnel sur le terrain dans lequel on peut recruter à tous les cinq ans des personnes expérimentées pour le recensement de la population. Compte tenu de ces facteurs, on a décidé de conserver l'organisation décentralisée.

3.3.2 Interviewers

Les nouveaux interviewers engagés en vue de l'Enquête sur la population active doivent effectuer cinq heures d'exercices et de lectures à domicile, suivies de trois jours de formation théorique; toutes ces heures sont rémunérées. Pendant leurs deux premières journées d'interviews du premier et du deuxième mois, les nouveaux interviewers sont observés par l'interviewer principal. En outre, les interviewers reçoivent régulièrement des documents à lire chez eux et des exercices à faire qui portent sur différents aspects des méthodes d'enquête. Ils peuvent également étudier chez eux les problèmes précis repérés au moment de la vérification des données au bureau central. Tous les interviewers reçoivent chaque année une formation théorique supplémentaire de trois jours. Dans le cas des enquêtes supplémentaires, la formation est donnée généralement à l'aide de documents à lire et d'exercices d'autofor- mation à faire chez soi. En ce qui regarde les enquêtes- entreprises, le nombre d'interviewers est beaucoup plus faible, soit 260 au total. La formation et l'encadrement sont les mêmes que dans le cas de l'Enquête sur la population active.

Dans une étude exhaustive sur la non-réponse, Gower (1979) a constaté que les taux de non-réponse variaient considérablement entre les interviewers. Fait particulièrement intéressant, environ 15% des interviewers enregistrés étaient régulièrement très peu de cas de non-réponse à l'ÉPA. Une étude portant sur des groupes d'interviewers de calibre supérieur et moyen tentera de découvrir en quoi ils diffèrent lorsqu'il s'agit de rejoindre les répondants et de les convaincre de participer à l'enquête. On envisagera ce dernier aspect du point de vue de la théorie de la conformité,

et les opérations de collecte de données jouent le rôle le plus direct et le plus important. Dans la présente section, nous examinerons les méthodes de collecte de données employées dans les enquêtes-entreprises et les enquêtes sociales ainsi que l'incidence qu'ont sur la non-réponse des facteurs tels que l'organisation, l'interviewer, le mode de collecte, la technologie, les méthodes de suivi et les incitatifs.

3.3.1 Organisation de la collecte de données

Les données des enquêtes-entreprises sont recueillies principalement au moyen d'enquêtes postales avec suivi téléphonique. Jusqu'au milieu des années 1980, la collecte et la vérification des données produites par les enquêtes-entreprises s'effectuaient principalement dans les divisions spécialisées de Statistique Canada, au bureau central. Par conséquent, plus de 70% du personnel de ces divisions était affecté au traitement des données d'enquête. Dans le cas de nombreuses enquêtes-entreprises, il incomrait aux bureaux régionaux de recueillir des données sur les non-répondants. Au milieu des années 1980, on s'est rendu compte qu'on pouvait mieux utiliser les ressources du bureau central et des bureaux régionaux en modifiant l'organisation de la collecte de données. Une division du bureau central a donc été chargée de la collecte et de la saisie de données annuelles, tandis que la collecte de données d'enquêtes intra-annuelles a été confiée aux bureaux régionaux. Les avantages de cette réorganisation sont les suivants: i) on peut utiliser plus efficacement les ressources opérationnelles; ii) on obtient une meilleure répartition des ressources entre le bureau central et les bureaux régionaux; iii) la collecte de données, activité de plus en plus complexe, peut être effectuée par des groupes spécialisés à cette fin et peut favoriser l'application d'innovations techniques et une plus grande intégration des méthodes de collecte; iv) les bureaux régionaux peuvent établir des contacts plus directs avec les répondants éventuels en raison de leur proximité géographique, et v) les bureaux régionaux peuvent offrir aux utilisateurs des services permettant de réhausser la présence de Statistique Canada auprès des unités répondantes éventuelles. Toutes ces mesures ont permis de réduire les taux de non-réponse. Les données des enquêtes sociales sont recueillies par une combinaison d'interviews directes et d'interviews téléphoniques. L'Enquête mensuelle sur la population active et la plupart des autres enquêtes sociales effectuées par Statistique Canada font appel à un millier d'interviewers dispersés dans tout le pays. Les interviewers effectuent à la fois des interviews téléphoniques depuis leur domicile et des interviews directes. Leur travail est surveillé par une centaine d'interviewers principaux. Des chargés de projet en poste dans chacun des bureaux régionaux de Statistique Canada supervisent le travail de quatre interviewers principaux. Dans le cas de l'EPA, les chargés de projet et les interviewers principaux reçoivent chaque mois des rapports sur le rendement des interviewers dont ils surveillent le travail. Ces rapports contiennent des mesures telles que les taux de rejet à la vérification, les taux de non-réponse et les coûts. La réaction permanente qu'impliquent ces rapports permet d'améliorer les procédures de collecte de

Si tous les aspects du plan de sondage peuvent avoir une incidence sur les taux de réponse enregistrés, les méthodes

3.3 Méthodes de collecte de données

déclaration.

d'une enquête à l'autre afin d'alléger le fardeau de que les échantillons de logements ne soient pas les mêmes interviewers à plus d'une enquête. On verra aussi à ce titillonnage (UPF) pour permettre d'affecter les mêmes avec chevauchement de certaines unités primaires d'échantillonnage également un échantillonnage coordonné, semblables ainsi qu'une stratification polyvalente. Elle une base de sondage commune et des plans d'échantillonnage éléments, l'orientation générale des enquêtes comprendra l'EPA, mais aussi de celles de ces enquêtes. Entre autres l'EPA tiendra compte non seulement des exigences de du revenu et une enquête sur la santé. Le remaniement de une enquête longitudinale sur la dynamique du travail et doivent être mises en route à la mi-décennie, notamment ménages. Plusieurs nouvelles enquêtes sociales périodiques consiste à définir un cadre général pour les enquêtes sur la population active, qui sera mis en oeuvre en 1995-1996, L'un des objectifs du remaniement de l'Enquête sur la choisies pour cette enquête.

compris dans le sous-ensemble attribué à une enquête sont (0, 1) et toutes les unités dont le nombre aléatoire est à différentes enquêtes des sous-ensembles de l'intervalle 0 et 1, à chaque unité de la population. On attribue alors attribuer un nombre aléatoire permanent, compris entre l'échantillonnage synchronisé, méthode qui consiste à chemement entre les enquêtes. À cette fin, on a recours à l'agriculture consiste à réduire au maximum le chevauchement entre les enquêtes sur les entreprises et sur unités visées par les enquêtes sur les petites entreprises de réduire le fardeau de déclaration des petites lement du 6^e de l'échantillon à chaque mois). Un autre mois), et l'Enquête sur la population active (avec renouvellement d'environ le 24^e des petites unités à chaque mensuelle sur le commerce de gros et de détail (avec renouvellement d'environ le 12^e des unités de travail (avec renouvellement d'environ le 12^e des unités sont l'Enquête sur l'emploi, la rémunération et les heures minimum. Les enquêtes qui utilisent un plan de ce genre nouveau pour la même enquête pendant un délai donné, au terme de laquelle elle ne peut être choisie de unité demeure dans l'échantillon pendant une période de déclaration. Selon les plans de renouvellement, une donnerait lieu à une répartition inacceptable du fardeau du changement, et l'absence de tout renouvellement, qui très coûteux et qui aboutit à des estimations insatisfaisantes d'un compromis entre un renouvellement complet, qui est d'échantillonnage selon un pourcentage donné; il s'agit choisit de procéder à un renouvellement partiel des unités préparer les nouvelles unités à fournir des données. On mentaire donnée aux interviewers et des difficultés pour supplémentaires de l'échantillon, de la formation supplémentaire le coût de l'enquête en raison des mises à jour d'échantillonnage. Toutefois, le renouvellement des unités petites unités en renouvelant périodiquement les unités Il est possible de réduire le fardeau de déclaration des

téléphoniques les numéros de téléphone des ménages actuellement ou récemment visés par l'EPA ou par d'autres enquêtes qui utilisent la base de sondage aréolaire. L'EPA est en train d'être remaniée. Il est question de faire appel à un registre d'adresses à titre d'échantillonnage sur liste dans les zones urbaines. On a établi un registre d'adresses de logements résidentiels afin d'améliorer la couverture du recensement de 1991; ce registre est mis à jour en fonction de l'énumération des logements tirée du recensement (Swain *et al.* 1992). On étudie actuellement la possibilité de mettre à jour continuellement ce registre à l'aide de dossiers administratifs ou de renseignements fournis par le service postal et de faire de ce registre la base de sondage des enquêtes sociales. Une base de sondage fondée sur un registre d'adresses devrait faciliter les opérations sur le terrain et réduire la non-réponse. Pour communiquer avec les ménages, les interviewers disposeront des numéros de téléphone des logements dans une proportion pouvant atteindre 70%. Grâce à sa mise à jour régulière, l'échantillon pourra être conçu de façon à ne pas alourdir la charge de travail des interviewers, sans que l'on doive recourir à des mesures telles que le sous-échantillonnage comme dans le cas de la base de sondage aréolaire. De plus, dans le cadre du remaniement, des mécanismes seront mis en place, tant pour la base de sondage aréolaire que pour la nomenclature, afin de retrouver tous les logements choisis aux fins des enquêtes de Statistique Canada.

3.2 Plan d'échantillonnage

On détermine la taille de l'échantillon d'une enquête en tenant compte des budgets, des objectifs de l'enquête et du degré de fiabilité recherché à l'égard des variables clés des principaux domaines visés. Le plan de sondage et la taille de l'échantillon global doivent être tels qu'il soit possible d'exercer un suivi des unités non répondantes. À la section 4, nous illustrerons ce point à l'égard des enquêtes mensuelles de Statistique Canada sur le commerce de gros et de détail, qui ont été remaniées récemment. Les enquêtes sur les entreprises et sur l'agriculture sont stratifiées selon un certain nombre de variables clés, dont la taille des unités. Comme la distribution des variables clés pour la population est très asymétrique, on se retrouve avec une strate à tirage complet et un certain nombre de strates à tirage partiel lorsque la taille sert de variable clé. Les unités de la strate à tirage complet ne peuvent être supprimées de l'échantillon par renouvellement, sauf si leur taille diminue avec le temps. Les plans d'échantillon-nage optimaux qui réduisent au maximum la taille de l'échantillon global pour des degrés de fiabilité donnés peuvent nécessiter un trop grand nombre d'unités dans la strate à tirage complet. Afin de réduire au maximum le fardeau des répondants, certaines enquêtes limitent le nombre d'unités à tirage complet, par exemple l'Enquête nationale sur les fermes (Julien et Maranda 1990). Dans le cas des répondants des grandes unités, on étudie également la possibilité d'intégrer les questionnaires ou les procédures de collecte de données de plus d'une enquête. Ainsi, on n'aurait à recueillir que les données statistiques propres aux différentes enquêtes.

forme de questions supplémentaires posées aux répondants à l'EPA. Certaines enquêtes, en raison de la longueur de l'interview ou du caractère délicat du sujet, ne se prêtent pas à la formule des suppléments; elles ont plutôt recours à des échantillons distincts de ménages tirés de la base de sondage de l'EPA.

L'EPA est surtout fondée sur une base de sondage aréolaire; le contact initial avec les ménages échantillonnés se fait en général par interview directe. L'efficacité de la base de sondage aréolaire décroît graduellement; les chiffres des logements des unités d'échantillonnage qui servent à déterminer les probabilités de sélection de ces unités et l'intervalle de sondage sont de moins en moins à jour. Il est donc plus difficile de planifier et d'assurer aux interviewers des charges de travail raisonnables. Le principal moyen de tenir à jour la base de sondage aréolaire consiste à remanier l'échantillon à la suite de chaque recensement décennal de la population. On a déjà pris d'autres mesures, notamment la mise à jour ponctuelle de la base de sondage, limitée aux secteurs à forte croissance repérés à la suite du recensement mi-décennal. Dans le cadre du remaniement de 1981, on a par ailleurs créé des 'strates tampons' à la périphérie des grands centres urbains. De conception simple, ces strates peuvent être mises à jour rapidement sans incidence sur le reste de la base de sondage au cas où le centre urbain empiéterait sur la zone tampon. Pour éviter que la charge de travail des interviewers devienne trop lourde lorsque des unités à forte croissance entrent dans l'échantillon, on procède à un sous-échantillonnage. Dans les cas de croissance extrême, on a recours à un sous-échantillonnage aréolaire, l'unité spatiale étant subdivisée en nouvelles unités dont on prélève un sous-échantillon. Si la croissance n'est pas trop forte, on conserve l'unité d'échantillonnage initiale. On modifie le taux d'échantillonnage pour réduire le nombre de logements prélevés de façon à ne pas alourdir la charge de travail de l'interviewer. Outre la base de sondage aréolaire, l'EPA comprend un échantillonnage sur liste des immeubles d'appartements des grandes villes. Cette liste est tenue à jour à l'aide de renseignements sur les permis de construire. Pour faciliter le contact avec les logements de l'échantillon d'appartements, on obtient, dans la mesure du possible, les numéros de téléphone en confrontant les adresses avec les dossiers de la compagnie de téléphone. Il s'est avéré utile de fournir ainsi les numéros de téléphone aux interviewers, car ils disposent alors d'un moyen supplémentaire pour contacter des logements échantillonnés qui sont difficiles d'accès en raison de la présence d'un système de sécurité ou dont les occupants sont rarement chez eux. Depuis l'adoption de cette méthode, même si le taux de non-réponse de la base des immeubles d'appartements demeure plus élevé que celui de la base aréolaire, l'écart s'est réduit, passant de 8,6% à 6,2%. Au lieu d'une base de sondage aréolaire, quelques enquêtes sociales utilisent une base de sondage téléphonique. L'échantillonnage est fondé sur un système d'appel aléatoire à partir de 'banques' de numéros résidentiels valides. Ces banques sont mises à jour à l'aide de dossiers achetés auprès des compagnies de téléphone. Afin d'alléger le fardeau des répondants, on exclut des enquêtes

Trois des enquêtes sociales présentent les taux de non-réponse les plus élevés, attribuables aux facteurs suivants: i) le sujet délicat que constitue le revenu dans le cas de l'Enquête sur les finances des consommateurs, ii) le fardeau des répondants, en raison de la longueur de l'interview pour l'Enquête sur les dépenses des familles, et iii) l'inexistence des interviewers, la méthodologie de l'enquête par téléphone et les déclarations sans personne interposée dans le cas de l'Enquête sociale générale. Le taux de non-réponse est très faible dans le cas de l'Enquête sur la population active (EPA) car il s'agit d'une enquête de premier plan qui existe depuis longtemps et dans laquelle de nombreuses mesures sont prises pour réduire le taux de non-réponse. Les taux de non-réponse sont faibles dans le cas des enquêtes-entreprises du tableau 1. Ce résultat est attribuable aux divers moyens mis en oeuvre à cette fin au cours du récent programme de remaniement des enquêtes-entreprises.

3. FACTEURS AVANT UNE INCIDENCE SUR LA NON-RÉPONSE

Plusieurs aspects des enquêtes ont une incidence sur les réponses et sur les non-réponses. Dans la présente section, nous commencerons par examiner brièvement l'influence de la base de sondage et du plan d'échantillonnage. Nous poursuivons par un examen approfondi de la collecte de données aux chapitres de l'organisation, de la formation des interviewers, de la technologie, du mode de collecte initiale et de la conception des questionnaires. Nous nous pencherons aussi sur les méthodes de suivi pour la non-réponse et les données rejetées à la vérification, et sur l'utilisation de données administratives en remplacement de la collecte directe.

3.1 Base de sondage

Le double compte ou le surdénombrement dans une base de sondage peuvent constituer des irritants et donner lieu à des non-réponses lorsqu'on ne prend pas de mesures pour assurer un dénombrement sans double compte ou si ces mesures ne s'avèrent pas toujours efficaces. Dans le cas des enquêtes-entreprises, il est essentiel de disposer de renseignements précis sur la classification si l'enquête est propre à un secteur d'activité ou si elle utilise des questionnaires propres à un secteur d'activité. Par exemple, une entreprise échantillonnée qui reçoit un questionnaire ne correspondant pas à son activité industrielle n'y répondra probablement pas. Il est nécessaire de disposer de renseignements précis sur la couverture d'entreprises à structure complexe pour fournir aux répondants une description précise des renseignements d'ordre géographique ou industriel voulus. De même, on a besoin de renseignements sur les personnes-ressources au sein de l'entreprise pour établir avec le répondant des modalités de déclaration efficaces. Si ces renseignements sont inexacts, l'obtention des données demandées s'en trouve retardée. D'autre part, si la description de la couverture est inexacte, le répondant risque de fournir des données non pertinentes ou incomplètes.

Les échantillons des enquêtes-entreprises de Statistique Canada sont tirés d'un fichier appelé Registre des entreprises. Ce fichier est une nomenclature contenant des renseignements pertinents qui permettent de former des échantillons d'entreprises répondantes et de communiquer avec ces dernières. On l'a remanié récemment à l'aide d'un modèle exhaustif qui reflète la complexité concrète des entreprises répondantes. Les procédés intégrés au Registre des entreprises réduisent au maximum l'incidence des causes de non-réponse mentionnées plus haut. Les doubles comptes sont maintenus au minimum grâce au lien établi continuellement entre les modifications et les unités qui figurent dans le Registre des entreprises. Ces modifications comprennent la création, la fusion, la séparation et l'absorption d'entreprises répondantes. Il y a plusieurs moyens de vérifier si des changements de structure sont survenus dans les grandes entreprises, notamment par différentes sources administratives et des enquêtes de rétroaction. L'observation d'un changement déclenche l'établissement d'un "profil": on communique avec l'entreprise pour redéfinir sa structure. En l'absence de signaux, on établit périodiquement le profil des structures à un rythme déterminé par leur importance et leur propension au changement. L'établissement de profils permet de recueillir les renseignements nécessaires à la mise à jour du modèle. Colledge (1989) traite plus en détail les mesures à mettre en oeuvre à cet égard. La source des mises à jour est une combinaison de mises à jour administratives, de mises à jour de profils et de réponses obtenues dans des enquêtes de rétroaction. À l'égard de chaque unité d'échantillonnage, on tient à jour les renseignements sur les personnes-ressources, la couverture et le type de questionnaire en créant et en mettant à jour un système de collecte informatisée de données pour les entreprises échantillonnées pour chaque enquête visée. L'établissement et la mise à jour des unités de collecte s'effectuent automatiquement selon des règles bien définies qui varient d'une enquête à l'autre. Le type de questionnaire tient compte de facteurs tels que la périodicité de la collecte de données, la classification industrielle, des critères saisonniers dans le cas d'enquêtes intra-annuelles, ainsi que la fin d'exercice dans le cas d'enquêtes annuelles. La mise à jour automatique de ces unités de collecte se fait sur la base d'une série de modifications entrées dans le Registre des entreprises. Ces modifications portent sur l'état des activités (entreprise active, inactive, saisonnière), le nom, l'adresse, le numéro de téléphone et la structure de l'unité enquêtée.

La pertinence de la base de sondage joue le même rôle pour réduire le taux de non-réponse dans les enquêtes sociales. Allée au plan d'échantillonnage et aux méthodes de collecte, la base de sondage est importante car elle assure aux interviewers une charge de travail raisonnable, elle fournit des renseignements qui faciliteront les contacts avec les répondants et elle élimine les chevauchements indésirés de l'échantillon d'une enquête à l'autre. L'Enquête sur la population active (EPA) est la principale enquête sociale dont l'exécution repose sur un sondage aréolaire. À l'heure actuelle, la plupart des autres enquêtes sociales sont des suppléments de l'EPA et prennent la

sociales, il s'agit des logements dont les occupants étaient provisoirement absents et des ménages dont aucun membre n'était présent lorsque l'interviewer est passé ou a appelé. On détermine si ces logements sont occupés ou non en observant les lieux ou, le cas échéant, en s'adressant au concierge de l'immeuble. Dans le cas des enquêtes-entreprises, il s'agit de répondants qu'on ne peut rejoindre par téléphone et de non-répondants dans le champ de l'enquête qui ont reçu un envoi postal mais qu'on n'a pas contactés lors d'un suivi des cas de non-réponse. Le **taux d'unités non contactées** se définit comme le *rapport du nombre d'unités non contactées et de cas non résolus au nombre d'unités dans le champ de l'enquête et de cas non résolus*. Les **unités de non-réponse résiduelles** sont les unités qui n'ont pas répondu en raison de circonstances particulières (par exemple, problèmes de langue ou incapacibilité) ainsi que les répondants qui n'ont pas fourni de renseignements utilisables. Parmi les circonstances particulières, mentionnons également la non-interview d'unités dans le champ de l'enquête pour éviter que les mêmes unités ne se retrouvent dans des échantillons d'enquêtes différentes et alléger ainsi le fardeau des répondants. Bien que ces dernières unités diffèrent des autres cas de non-réponse parce qu'on ne tente pas de les interviewer, on se doit de les considérer comme des non-répondants pour établir

Tableau 1

Répartition des taux de réponse pour certaines enquêtes à l'étape de la collecte de données
(taux en pourcentage)

CALCUL	BDF	EAIM	ESG	EFC	BPA	BCD
--------	-----	------	-----	-----	-----	-----

Taux de cas résolus	(2)/(1)	100.0	100.0	98.1	100.0	100.0	95.8
Taux d'unités dans le champ	(4)/(2)	92.0	95.3	51.2	86.3	85.1	97.0
Taux de réponse	(6)/[(3) + (4)]	72.9	92.8	75.9	73.9	94.4	94.0
Taux de réponse après refus	(11)/[(11) + (13)]	S.O.	S.O.	26.7	S.O.	S.O.	S.O.
Taux de non-réponse	[(7) + (3)] / [(3) + (4)]	27.1	7.2	24.1	26.1	5.6	6.0
Taux de refus	(13)/(4)	16.2	7.2	13.2	23.7	1.5	1.7
Taux d'unités non contactées	[(14) + (3)] / [(3) + (4)]	5.1	0.0	5.9	2.3	3.6	4.3
Taux de non-réponse résiduelle	(15)/(4)	5.8	0.0	5.8	0.0	0.4	0.0
Taux d'unités hors du champ	(5)/(2)	8.0	4.7	48.8	13.7	14.9	3.0
Taux d'unités inexistantes	(8)/(2)	0.8	2.5	0.0	0.3	0.3	2.3
Taux d'unités provisoirement hors du champ	(9)/(2)	7.1	1.2	0.0	13.4	14.6	0.5
Taux d'unités hors du champ en permanence	(10)/(2)	0.0	1.0	48.8	0.0	0.0	0.3

EDF: Enquête sur les dépenses des familles (1990).
EAIM: Enquête annuelle sur les industries manufacturières (1989).
ESG: Enquête sociale générale, 5e cycle (janvier-mars 1990).
EPA: Enquête sur la population active (1990).
EFC: Enquête sur les finances des consommateurs (1991).
BCD: Enquête sur le commerce de détail (décembre 1990).

Les unités dans le champ de l'enquête (case 4) se répartissent en unités répondantes (case 6) et non répondantes (case 7). Les **unités répondantes** sont les unités dans le champ de l'enquête qui ont répondu à la date limite de collecte de données et qui ont fourni des "renseignements utilisables". La notion de "renseignements utilisables" s'applique aux répondants qui ne fournissent que des renseignements partiels. Il est nécessaire d'établir un seul de réponses au questionnaire, au-dessous duquel on considère une unité comme un non-répondant. On peut définir le **taux de réponse** de diverses façons, selon l'analyse envisagée. Selon notre définition, il s'agit du rapport du nombre d'*unités répondantes au nombre d'unités dans le champ de l'enquête et de cas non résolus*. Ce rapport est une mesure prudente de la qualité de la base de sondage et de la méthode de collecte de données puisque certains cas non résolus peuvent représenter des unités hors du champ de l'enquête. Une autre définition du taux de réponse pourrait être le rapport du nombre d'*unités répondantes au nombre d'unités dans le champ de l'enquête*. Dans ce dernier cas, il s'agit d'un taux de réponse conditionnel étant donné l'état des unités qui composent l'échantillon et ce taux sert à mesurer l'efficacité de la méthode de collecte de données seulement. Les **unités non répondantes** (case 7) englobent le reste des unités dans le champ de l'enquête. Le **taux de non-réponse** se définit comme le complément du taux de réponse, soit le rapport du nombre d'*unités non répondantes et de cas non résolus au nombre d'unités dans le champ de l'enquête et de cas non résolus*. D'autres définitions excluent les cas non résolus du numérateur et du dénominateur ou en font deux sous-groupes : nombre estimé d'*unités dans le champ de l'enquête et nombre estimé d'unités hors du champ de l'enquête* et le taux d'*unités hors du champ de l'enquête* et le taux d'*unités hors du champ de l'enquête* en permanence.

À moins de vérifier auprès de la compagnie de téléphone l'état de chaque numéro où personne ne répond, il est impossible d'établir s'il s'agit d'un numéro valide. De même, dans le cas d'une enquête avec envoi postal, faute d'un suivi des unités qui ne retournent pas le questionnaire, on ne peut savoir lesquelles sont hors du champ de l'enquête (ex. : l'unité n'existe plus ou, si elle existe toujours, elle est hors du champ de l'enquête) ni celles qui sont dans le champ de l'enquête et qui auraient dû répondre. Les cas **non résolus** représentent les unités dont on ne peut établir l'état à la fin de l'étape de la collecte de données. On peut diviser ce groupe d'unités en deux sous-groupes – **nombre estimé d'unités dans le champ de l'enquête** et **nombre estimé d'unités hors du champ de l'enquête** – selon les mêmes rapports que ceux utilisés pour les cas résolus par exemple. On peut donc définir ainsi le **taux de cas résolus : rapport du nombre de cas résolus au nombre total d'unités**. Les deux sous-groupes des cas résolus, soit les unités dans le champ de l'enquête (case 4) et les unités hors du champ de l'enquête (case 5) sont à l'origine de deux taux complémentaires : le **taux d'unités dans le champ de l'enquête**, soit le rapport du nombre d'*unités dans le champ de l'enquête et de cas résolus*, et son complément, le **taux d'unités hors du champ de l'enquête**. Les unités hors du champ de l'enquête (case 5) se répartissent jusqu'en trois catégories, dont certaines ne s'appliquent pas nécessairement à une enquête donnée. Il s'agit des unités inexistantes (case 8), des unités provisoirement hors du champ de l'enquête (case 9) et des unités hors du champ de l'enquête en permanence (case 10). Les **unités inexistantes** comprennent les entreprises disparues, soit celles qui ont fermé leurs portes, et les logements démolis. Dans le cas d'enquêtes périodiques, dès qu'il est établi qu'une unité est inexistante, celle-ci est exclue de la collecte ultérieure de données. Les **unités provisoirement hors du champ de l'enquête** sont les unités qui étaient hors du champ d'observation au moment de l'enquête, mais qui pourraient s'y retrouver ultérieurement. On peut donc trouver des unités provisoirement hors du champ de l'enquête même dans des enquêtes uniques. Lorsqu'il s'agit d'enquêtes périodiques, il est nécessaire de recom-muniquer périodiquement avec les unités provisoirement hors du champ de l'enquête au cas où leur état aurait changé. Citons, par exemple, les entreprises inactives en raison de facteurs saisonniers, les logements saisonniers dont les occupants possèdent une résidence principale ailleurs ou les logements vacants. Les **unités hors du champ de l'enquête en permanence** sont des unités qui n'ont pas leur place dans la base de sondage à cause de modifications qui ont été apportées à la classification depuis la dernière mise à jour de la base. On peut déceler et éliminer ces cas à la première étape de la réception des réponses. Le taux d'unités hors du champ de l'enquête se répartit en trois taux constitutifs : le **taux d'unités inexistantes**, soit le rapport du nombre d'*unités inexistantes au nombre de cas résolus et, définis de la même façon, le taux d'unités provisoirement hors du champ de l'enquête et le taux d'unités hors du champ de l'enquête en permanence.*

des données sont recueillies. Nous ne traitons pas la non-réponse partielle, c'est-à-dire le cas où l'unité répondante fournit de l'information utilisable pour certaines questions mais non pour d'autres. Nous commencerons par établir un cadre conceptuel permettant de définir la réponse et la non-réponse et convenant à la fois aux enquêtes-entreprises et aux enquêtes sociales. Nous aborderons ensuite les causes générales de la non-réponse et les façons de réduire le nombre de non-réponses. Enfin, nous examinerons les taux de non-réponse enregistrés dans deux grandes enquêtes de Statistique Canada.

2. DÉFINITION DES TAUX DE NON-RÉPONSE

Les taux de non-réponse et leur contrepartie, les taux de réponse, se définissent comme des rapports de variables qui représentent une catégorie donnée de réponse ou de non-réponse dans un certain domaine. La variable d'importance peut être un simple chiffre ou une donnée pondérée par un facteur quelconque, par exemple le poids de l'unité dans l'échantillon ou la part prévue de l'unité dans l'estimation d'une statistique importante dégagée par l'enquête. La figure 1 représente un cadre conceptuel élaboré par Drew et Gray (1991) pour diviser les unités

d'échantillonnage d'une enquête en répondants, en non-répondants et en unités hors du champ de l'enquête. L'organigramme est semblable à celui qu'avaient déjà proposé Platek et Gray (1986). Statistique Canada a vérifié ce cadre en l'appliquant à plusieurs enquêtes-entreprises et enquêtes sociales et l'a adopté comme norme en matière de collecte et de communication de renseignements sur la non-réponse. À compter de l'année de référence 1993, il faudra compter sur plusieurs grandes enquêtes pour produire des renseignements détaillés sur la non-réponse en se servant de données dans laquelle seront versés les taux de non-réponse et cette base permettra de surveiller la tendance de la non-réponse dans l'ensemble de l'organisme et d'en faire l'analyse.

Commentons par le nombre total d'unités (pondérées ou non). Il s'agit des unités susceptibles d'être visées par l'enquête avant le commencement de celle-ci. Le total (case 1 de la figure 1) se divise en deux grandes catégories: les cas résolus (case 2) et les cas non résolus (case 3). Les cas résolus représentent les unités dont on sait si elles appartiennent ou non à l'univers cible à la date limite de collecte de données. Pour certaines enquêtes, toutes les unités sont des cas résolus. Pour d'autres, il est impossible ou peu pratique de résoudre tous les cas. Dans une enquête téléphonique, par exemple, on trouve des numéros où le téléphone

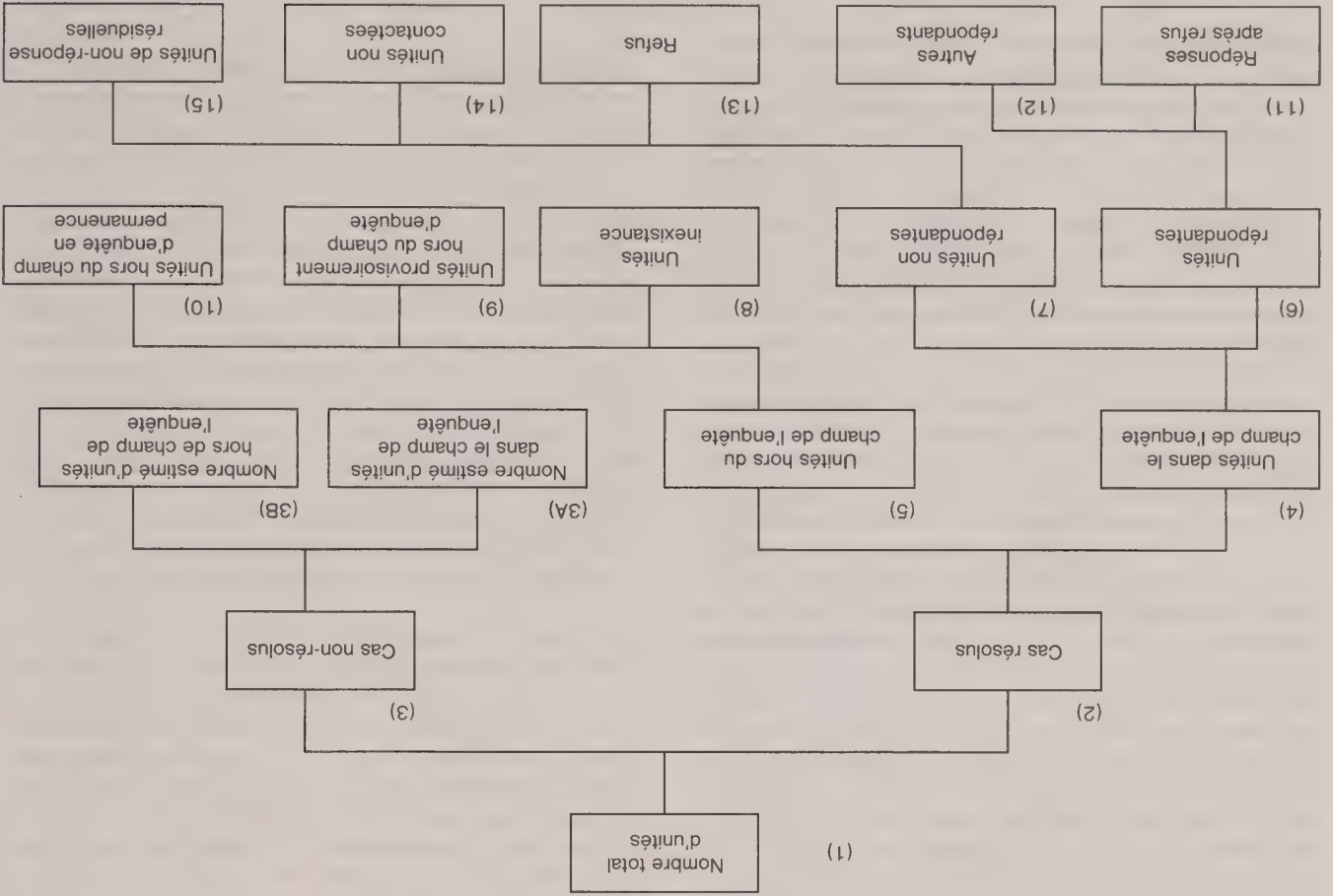


Figure 1. Répartition des répondants et des non-répondants à l'étape de la collecte de données

Cadre pour l'évaluation et la réduction de la non-réponse dans les enquêtes

MICHAEL A. HIDIROGLOU, J. DOUGLAS DREW
et GERALD B. GRAY¹

RÉSUMÉ

Nous traiterons de la nécessité d'établir des normes en matière de collecte et de communication de renseignements sur les non-réponses aux enquêtes effectuées par un organisme de statistique, puis nous décrirons les normes adoptées par Statistique Canada. Nous présenterons ensuite les mesures prises à différentes étapes de la conception des enquêtes à Statistique Canada dans le but de réduire la non-réponse. Nous illustrerons ces points en examinant les taux de non-réponse pour deux grandes enquêtes de Statistique Canada.

MOTS CLÉS: Taux de non-réponse; incitatifs; suivis; collecte de données.

1. INTRODUCTION

Les organismes nationaux comme Statistique Canada effectuent chaque année un grand nombre d'enquêtes très variées. Ces enquêtes varient selon le domaine, les unités de réponse, la périodicité, le plan de sondage et les méthodes de collecte de données. Elles enregistrent également des taux de non-réponse différents. Il est nécessaire d'établir, pour l'ensemble de l'organisme, des normes en matière de collecte et de communication de renseignements sur les taux de réponse et de non-réponse. Si ces normes sont suffisamment souples pour répondre aux exigences des diverses enquêtes effectuées par l'organisme, on peut logiquement introduire des définitions normalisées. Il faut toutefois faire la distinction entre des définitions normalisées et ce que l'on peut juger acceptable comme niveau pour les différentes composantes de la non-réponse dans les enquêtes. Cet article porte sur le premier point et non sur le second.

On observe des différences majeures entre des enquêtes qui enregistrent des taux de non-réponse différents; par exemple, les problèmes d'absence de données ne sont pas les mêmes dans les enquêtes longitudinales et les enquêtes transversales. Les définitions normalisées peuvent susciter un vocabulaire commun qui permettra de typifier ces différences et de mieux les comprendre. Un vocabulaire commun facilite l'analyse continue des tendances de la non-réponse. L'information sur les réponses et les non-réponses à une enquête peut servir à de nombreuses fins: renseigner les utilisateurs sur la possibilité d'un biais de non-réponse, relever les lacunes à corriger dans des éditions ultérieures de l'enquête, mesurer la couverture de la base de sondage, élaborer des méthodes visant à compenser et à réduire la non-réponse. En outre, cette information est un élément important dans la conception des enquêtes et l'élaboration des méthodes de collecte de données, dans l'évaluation de la qualité des données et les opérations des enquêtes.

On peut définir différemment les taux de non-réponse selon qu'ils servent à évaluer des activités d'échantillonnage ou de collecte de données ou à analyser des données publiées. En matière d'échantillonnage, par exemple, l'unité par rapport à laquelle on mesure le taux de non-réponse sera l'unité d'échantillonnage. De même, en matière de collecte de données, l'unité de mesure servant à calculer le taux de non-réponse sera fondée sur l'unité de réponse. Souignons que dans le cas des enquêtes-entreprises, il y a rarement correspondance exacte entre les unités d'échantillonnage et les unités de réponse (par ex., l'unité échantillonnée peut être le siège social tandis que l'unité de réponse sera la succursale). Pour ce qui est des données publiées, la mesure du taux de non-réponse peut être représentée par des mesures de taille pondérées ou des variables clés pondérées en vue d'estimer la part des non-répondants dans les agrégats clés. Dans le cas des enquêtes-entreprises, ces mesures peuvent être importantes en raison de la présence de populations asymétriques, dans lesquelles un petit nombre d'unités représentent un pourcentage disproportionné de l'estimation. Les taux devraient être ventilés en fonction de divers critères préétablis, pris isolément ou combinés: zones géographiques, secteurs d'activité, taille. Dans la mesure du possible, on doit également connaître les motifs de non-réponse (par ex., impossibilité de communiquer avec le répondant, refus, etc.), qui peuvent servir à diagnostiquer les causes de la non-réponse. Si les données sont recueillies par des interviewers en poste dans les bureaux régionaux de tout le pays, des statistiques comme les taux de non-réponse par interviewer au sein de chaque bureau régional et les taux de non-réponse par bureau régional peuvent servir à mesurer le rendement opérationnel. Les taux de non-réponse à une question peuvent servir à mettre en évidence les questions qui mériteraient d'être reformulées ou redéfinies en fonction de la disponibilité des données. Cet article porte sur la non-réponse totale, c'est-à-dire la non-réponse enregistrée au niveau de l'unité sur laquelle

¹ Michael A. Hidiroglou, Division des méthodes d'enquêtes-entreprises; J. Douglas Drew, Division des enquêtes des ménages; Gerald B. Gray, Division des méthodes d'enquêtes sociales, Statistique Canada.

SINGH, A.C., ARMSTRONG, J.B., et LEMAITRE, G.E. (1988). Statistical matching using log linear imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 672-677.

SINGH, A.C., MANTTEL, H., KINACK, M., et ROWE, G. (1990). On methods of statistical matching with and without auxiliary information: Some modifications and an empirical evaluation. Direction de la méthodologie, document de travail, SSMD, 90-016E. Statistique Canada.

U.S. DEPARTMENT OF COMMERCE (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, Washington, D.C.: Federal Committee on Statistical Methodology.

WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B., et ROWE, G. (1987). The social policy simulation database: an example of survey and administrative data integration. *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada, Ottawa, (Eds. J.W. Coombs et M.P. Singh), 201-229; une autre version publiée dans *Survey of Current Business* (1989), 69, 36-40.

des données réelles pour vérifier la solidité des résultats obtenus avec des données fictives et pour voir quel genre d'incidence peuvent avoir les biais contenus dans la distribution conjointe du fichier enrichi. On peut aussi chercher à savoir comment prendre en compte ces biais dans des inférences qui reposent sur le fichier enrichi, c'est-à-dire comment produire des mesures d'incertitude pour les estimations de paramètres tirées du fichier enrichi qui reflètent non seulement la variabilité dans ce fichier, mais aussi la variabilité inhérente à la procédure d'appariement. Bien que nous ne puissions répondre à cette question, il ne fait aucun doute que les méthodes d'appariement qui utilisent de l'information supplémentaire accroîtront l'utilité générée du fichier enrichi. Nous nous pencherons sur ces questions et sur d'autres questions connexes dans des études ultérieures.

REMERCIEMENTS

Les auteurs tiennent à remercier J. Armstrong, G. Gray, G. Hole, D. Royce et M. Wolfson pour leurs précieux commentaires. La recherche faite par le premier auteur a été rendue possible en partie grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada accordée à l'Université Carleton.

BIBLIOGRAPHIE

- ARMSTRONG, J. (1989). An evaluation of statistical matching methods. Direction de la méthodologie, document de travail, BSM-D, 90-003E. Statistique Canada.
- BARR, R.S., et TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Surveys. Rapport technique, U.S. Department of the Treasury, Office of Tax Analysis.
- BARR, R.S., et TURNER, J.S. (1990). Quality issues and evidence in statistical file merging. Dans *Data Quality Control: Theory and Pragmatics* (Eds. G.E. Liepins et V.R.R. Uppluri). New York: Marcel Dekker, 245-313.
- BARR, R.S., STEWART, W.H., et TURNER, J.S. (1981). An empirical evaluation of statistical matching methodologies. Rapport technique, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.
- BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.
- BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.
- BUDD, E.C., et RADNER, D.B. (1969). The OBE size distribution series: methods and tentative results for 1964. *American Economic Review, Papers and Proceedings*, LIX, 435-449.
- COHEN, M.L. (1991). Statistical matching and microsimulation models. Dans *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Volume II, Technical Papers, (Eds. C.F. Citro et E.A. Hanushek). Washington, D.C.: National Academy Press, 62-85.
- FELLEG, I.P. (1977). Discussion paper. *Proceedings of the Section on Social Statistics, American Statistical Association*, 762-764.
- FORD, B.L. (1983). An overview of hot-deck procedures. Dans *Incomplete Data in Sample Surveys*, (Vol. 2), (Eds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 185-207.
- KADANE, J.B. (1978). Some statistical problems in merging data files. Dans *1978 Compendium of Tax Research*. Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.
- KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.
- LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.
- PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. Dans *Microanalytic Simulation Models to Support Social and Financial Policy* (Eds. G.H. Orcutt, J. Merz et H. Quinke). Amsterdam: Elsevier Science.
- PAASS, G. (1989). Stochastic generation of a synthetic sample from marginal information. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 431-445.
- PAASS, G., et WAUSCHKUN, U. (1980). Experimentelle erprobung und vergleichende Bewertung statistischer Matchverfahren. Rapport interne, IPES.80.201, St. Augustin, *Gesellschaft für Mathematik und Datenverarbeitung*.
- PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or Domains). *Revue Internationale de Statistique*, 48, 3-18.
- RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91-102.
- RODGERS, W.L., et DEVOL, E. (1982). An evaluation of statistical matching. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-132.
- RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87-94.
- RUGGLES, N., RUGGLES, R., et WOLFF, E. (1977). Merging microdata: Rationale, practice and testing. *Annals of Economic and Social Measurement*, 6, 407-428.
- SCHUBERT, F.J. (1989). Comment on Wolfson et al. (1989). *Survey of Current Business*, 69, 40-41.
- SIMS, C.A. (1972). Comment on Okner (1972). *Annals of Economic and Social Measurement*, 1, 343-345.
- SIMS, C.A. (1978). Comment on Kadane (1978). Dans *1978 Compendium of Tax Research*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.
- SINGH, A.C. (1988). Log-linear imputation. Direction de la méthodologie, document de travail, SSM-D, 90-016E. Statistique Canada; publié aussi dans *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 118-132.

iiii) Comparaison de différentes versions de la méthode hot-deck

Dans toutes les méthodes d'appariement étudiées, sauf HOD et HOD.LOGLIN, la deuxième étape, qui consiste à déterminer la valeur Z_{match} , fait intervenir l'imputation hot-deck avec la distance (X, Z) . Pour les deux cas d'exception, on utilise la distance X . On peut aussi utiliser (pour d'autres méthodes que HOD et HOD.LOGLIN) la distance Z ou la distance (X, Y, Z) . Dans ce dernier cas, il faut voir tout d'abord à introduire la valeur X_{int} dans le fichier B. C'est ce qui a été fait dans l'étude de simulation originale, sauf que les résultats empiriques pertinents ne sont pas présentés ici. Les résultats ont montré que le choix de la distance change peu de choses; pour les méthodes REG et REG*, la distance (X, Z) donnait parfois des résultats supérieurs en ce qui a trait à la mesure AD-Cov. C'est ce qui nous a fait opter pour la distance (X, Z) pour les méthodes étudiées ici. Cependant, dans la pratique, il serait préférable d'utiliser la distance Z pour les méthodes hot-deck à cause de la simplicité des calculs. Par ailleurs, notons qu'en ce qui concerne les méthodes HOD et HOD.LOGLIN, on a la possibilité d'utiliser à l'étape 2 la version aléatoire de la méthode hot-deck ou la version fondée sur le rang, plutôt que celle fondée sur la distance X . Selon la version fondée sur le rang, les entregistremets des fichiers A et B sont ordonnés séparément en fonction de la valeur de X , puis ils sont appariés selon le rang. C'est la procédure qu'a proposée G. Rowe pour l'application relative à la BDSPS évoquée dans l'introduction. De toute évidence, cette méthode ne convient qu'aux variables X unidimensionnelles. Un avantage de la méthode fondée sur le rang est qu'un enregistrement du fichier B ne peut servir de donneur pour plus d'un entregistrement du fichier A. Les trois versions de hot-deck mentionnées ci-dessus étaient incluses dans l'étude de Monte Carlo mais seuls les résultats relatifs à la version fondée sur la distance X sont présentés ici. Notre étude a permis de constater que le choix de la version n'a pas beaucoup d'incidences en règle générale. Nous avons choisi la version fondée sur la distance X pour les méthodes HOD et HOD.LOGLIN parce qu'elle était conforme à la version de la hot-deck utilisée pour les autres méthodes (c.-à-d. celle fondée sur la distance). En pratique, la version léatoire serait la moins exigeante sur le plan du calcul; cependant, dans une application réelle, nous ne pourrions pas mesurer les inconvénients rattachés à l'utilisation de la méthode aléatoire par opposition à l'utilisation de la méthode fondée sur le rang ou la distance et pourtant, nous voudrions probablement disposer du plus de renseignements possible.

8. CONCLUSIONS

Cet article nous a permis de traiter le problème de l'utilisation d'informations supplémentaires pour l'appariement statistique. Nous avons vu que les deux principales méthodes d'appariement proposées antérieurement venaient de Rubin (1986) et de Paass (1986); des versions

de ces méthodes sont désignées par les sigles REG* et HOD* respectivement. Nous avons proposé des versions modifiées de ces méthodes, désignées par REG.LOGLIN* et HOD.LOGLIN*, en établissant des contraintes nominales à partir d'informations supplémentaires. Ces deux dernières versions se ramènent à REG.LOGLIN et à HOD.LOGLIN si l'on ne dispose que d'informations supplémentaires de type nominal. En l'absence d'imputation supplémentaire, on utilise les méthodes d'imputation habituelles, REG et HOD. Nous avons fait une étude empirique pour évaluer le rendement de huit méthodes d'appariement statistique par rapport à quatre mesures d'évaluation (deux au niveau individuel et deux au niveau global). L'étude a permis de constater que lorsqu'il n'y a pas d'information supplémentaire, la méthode HOD est préférable. Cependant, lorsqu'on dispose d'information supplémentaire, la situation est plus complexe. Si l'on juge importantes seulement les mesures d'évaluation au niveau individuel, alors la méthode REG* est recommandée. Si les mesures au niveau global sont elles aussi jugées importantes, la méthode HOD.LOGLIN* est recommandée si tant que l'information supplémentaire est non substitutive. Par ailleurs, la méthode HOD.LOGLIN est une bonne solution de compromis si le nombre de calculs à effectuer est une source de préoccupation majeure ou si l'on croit que l'information supplémentaire peut être substitutive. Si les mesures au niveau individuel sont moins importantes ou qu'elles ne présentent aucun intérêt (cela peut être fréquent parce que les données appariées sont souvent présentées en tableaux dans la pratique), alors HOD.LOGLIN sera la méthode recommandée. Compte tenu du rendement équivalent des trois versions de la méthode hot-deck (celles fondées sur la distance et le rang et la version aléatoire) pour les méthodes HOD et HOD.LOGLIN, on pourrait envisager d'utiliser la version aléatoire dans la pratique en raison de la simplicité des calculs.

Il convient de souligner que nous n'avons pas considéré la version entièrement itérative de la méthode de Paass. Il serait intéressant d'examiner dans des analyses futures le rendement de cette version. Une autre question qui mérite d'être étudiée est l'application de contraintes nominales avec de nombreuses variables. L'application de l'algorithme de "balayage" pourrait comporter un nombre exorbitant de calculs dans ce cas. À cet égard, les résultats de l'étude de Paass (1989) devraient être utiles. Dans l'étude que nous venons d'exposer, nous n'avons pas fait varier systématiquement le degré d'exactitude de la source de données supplémentaires à cause de limites imposées par le calcul; autrement dit, nous n'avons pas fait varier la taille du fichier C, ni d'ailleurs celle de fichiers A et B. Une question intéressante aurait été de savoir comment la taille de ces fichiers influe sur le rendement des diverses méthodes. Enfin, soulignons que même si les résultats de cette étude reposent sur des données fictives (ce qui était nécessaire pour créer des scénarios qui simulent des données réelles), nous croyons qu'ils seraient pertinents pour des applications concrètes. De toute évidence, il serait intéressant et utile de faire une étude de simulation avec

Les très fortes valeurs négatives observées aux extrêmes des graphiques sont associées à des intervalles ouverts, et il semble fort probable que si ces intervalles avaient été divisés en de plus petits intervalles, on aurait observé plusieurs valeurs négatives moindres aux extrêmes des graphiques, de sorte qu'on aurait pu déduire que les méthodes REG et REG* placent un trop grand nombre de valeurs Z au centre de la distribution par rapport aux extrêmes. La figure 1 permet de constater qu'il y a aussi une réduction à la moyenne dans le cas des méthodes REG, LOGLIN et REG.LOGLIN*. Toutefois, dans ce cas-ci la réduction est limitée par les contraintes nominales, de sorte que même si nous constatons que les extrêmes de la distribution de Z pour le fichier enrichi ne sont pas assez étendus, les valeurs Z ne sont pas concentrées au centre de la distribution mais s'agglomèrent autour des limites d'intervalle, qui ont l'effet d'un mur. On peut expliquer de la même manière les fortes valeurs positives observées de part et d'autre de la limite centrale si l'on tient compte du fait que ces graphiques représentent en réalité une moyenne des différences entre histogrammes pour 100 simulations indépendantes. Il semble raisonnable de penser que si nous devons étudier séparément chacune des 100 différences d'histogrammes, nous observerions une forte valeur positive tantôt juste à gauche de la limite centrale, tantôt juste à droite, mais jamais des deux côtés à la fois.

La figure 1 montre aussi que l'effet de réduction à la moyenne et l'effet des limites sont négligeables pour les méthodes de type hot-deck.

ii) Information supplémentaire (Y,Z) vs information supplémentaire (X,Y,Z)

Bien que cet article ne présente que les résultats fondés sur l'information supplémentaire (Y,Z), l'étude de simulation considérait aussi, comme nous l'avons dit dans la section 6, l'information supplémentaire (X,Y,Z). Nous avons constaté avec intérêt que dans le cas des méthodes HOD, LOGLIN et HOD, LOGLIN*, l'utilisation d'informations supplémentaires (Y,Z) produit, en règle générale, de meilleurs résultats au niveau global que l'utilisation d'informations (X,Y,Z). Cette constatation ne semble pas s'appliquer à la méthode HOD*. Ce phénomène s'explique probablement par la variabilité de l'estimation des effets des facteurs (X*,Y*,Z*) qui entrent dans les contraintes nominales, cette variabilité étant le fait d'un trop petit volume de données supplémentaires. Nous pouvons en conclure que les valeurs vraies étaient probablement voisines de zéro et que les assimiler à la valeur nulle donne de meilleurs résultats. C'est ce qui nous porte à croire qu'il faudrait tenir compte dans les analyses futures de l'effet de la taille de l'échantillon sur le rendement des méthodes d'appariement. Les considérations ci-dessus suggèrent aussi une nouvelle catégorie intéressante de méthodes qui combinerait l'information supplémentaire de micro-niveau (X,Y,Z) qui sert à déterminer les valeurs Z_{int} avec la distribution de variables nominales (Y*,Z*) tirée du fichier C qui sert à définir des contraintes. Cependant, ces méthodes n'ont pas été considérées dans la présente étude.

En ce qui concerne les méthodes hot-deck avec information supplémentaire, les deux méthodes assujetties à des contraintes nominales – notamment HOD, LOGLIN et HOD, LOGLIN* – ont un rendement très acceptable au niveau global (c.-à-d. par rapport à χ^2 transformé et au TRV) pour tous les types de population; voir figures 2 à 5. De façon générale, la méthode HOD, LOGLIN est supérieure à la méthode HOD, LOGLIN*. Considérons maintenant les mesures d'évaluation au niveau individuel. Pour des populations symétriques et dissymétriques (figures 2 et 3), les méthodes HOD* et HOD, LOGLIN* sont, en règle générale, aussi efficaces l'une que l'autre et sont un peu plus efficaces que HOD, LOGLIN mais un peu moins efficaces que REG*. Cependant, lorsqu'il s'agit d'information supplémentaire substitutive (figures 4 et 5), la méthode HOD, LOGLIN peut être soit plus efficace ou moins efficace que les méthodes HOD, LOGLIN* et REG* en ce qui concerne $AD-Cov$, et elle est souvent plus efficace lorsqu'il s'agit de données substitutives avec des contaminations log-normales. En outre, avec des données substitutives, HOD, LOGLIN tend à offrir assez de robustesse à l'égard des quatre mesures d'évaluation. Par conséquent, du point de vue du rendement global, nous pouvons recommander l'usage de la méthode HOD, LOGLIN* parmi toutes les méthodes hot-deck. Cependant, au point de vue pratique, il peut être préférable d'opter pour la méthode HOD, LOGLIN comme solution de compromis parce qu'elle est assez efficace au niveau individuel et extrêmement efficace au niveau global, qu'elle est beaucoup moins exigeante sur le plan du calcul et qu'elle montre de la robustesse à l'égard des données supplémentaires substitutives. En outre, HOD, LOGLIN ne nécessite pas d'information supplémentaire de micro-niveau.

7.3 Observations diverses

Dans cette sous-section, nous exposons quelques conclusions intéressantes qui sont ressorties de l'étude de Monte Carlo; les résultats empiriques relatifs à certaines de ces conclusions ne figurent pas ici mais sont présentés dans Singh *et al.* (1990).

i) Réduction à la moyenne

Une conclusion importante qui ressort de façon constante est que les méthodes d'appariement fondées sur la régression sont peu efficaces en ce qui concerne les mesures de type nominal. Ce manque d'efficacité peut s'expliquer par le phénomène de réduction à la moyenne, c'est-à-dire le phénomène selon lequel les valeurs Z appariées sont plus concentrées autour de leur moyenne que ne le sont les valeurs Z supprimées. C'est ce que l'on peut voir à la figure 1, qui montre la différence entre les histogrammes marginaux des valeurs Z appariées et des valeurs Z supprimées pour diverses méthodes d'appariement. Les écarts positifs que l'on observe près du centre pour REG et REG* indiquent que le fichier enrichi compte un plus grand nombre de valeurs Z dans cette région que le fichier duquel ont été supprimées les valeurs vraies de Z.

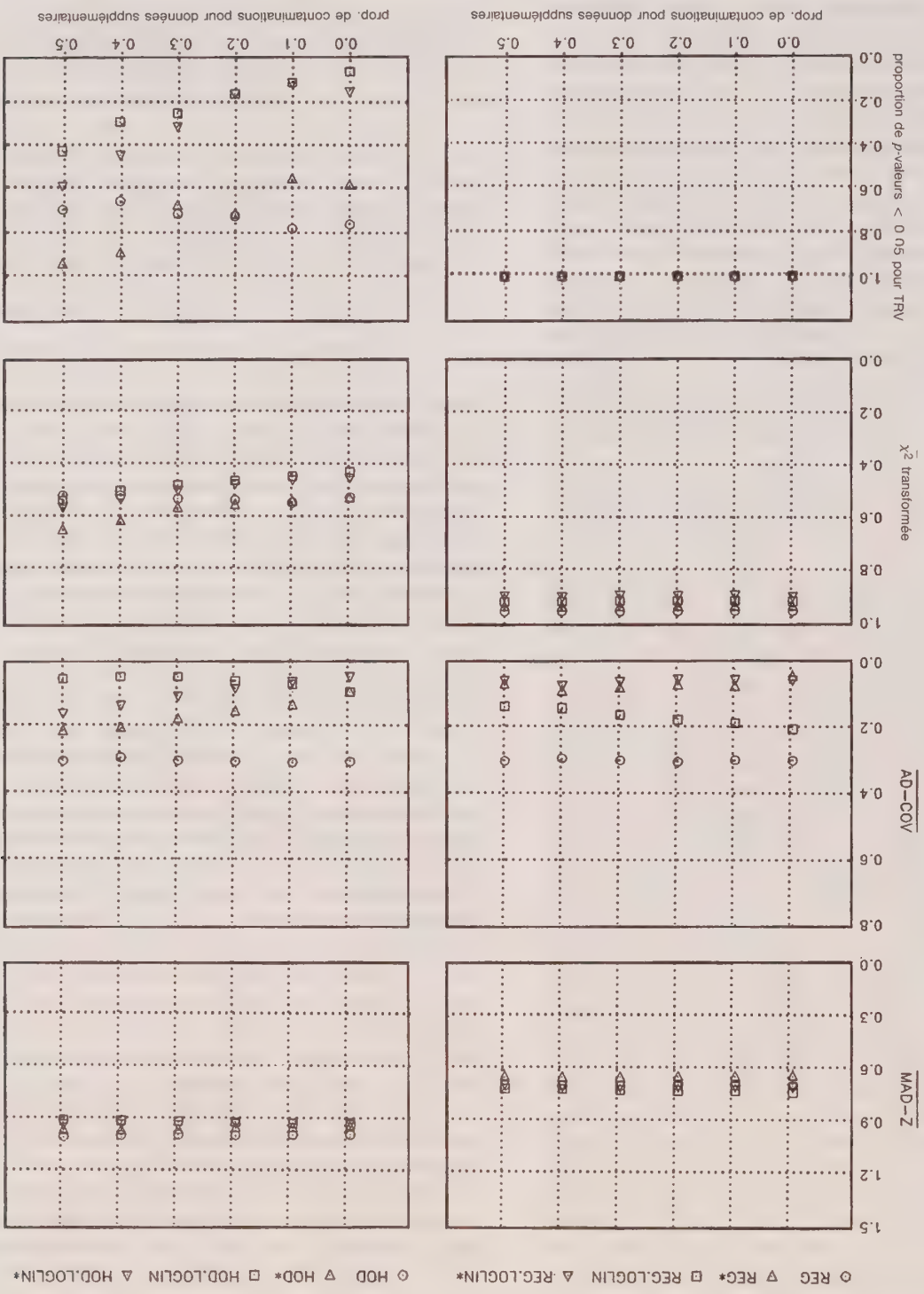


Figure 5. Comparaison de méthodes d'appariement statistique lorsque la proportion de contaminations log-normales pour le fichier C varie ($p_{Y,Z|X} = .4$ avant contamination et absence de contamination pour les fichiers A et B)

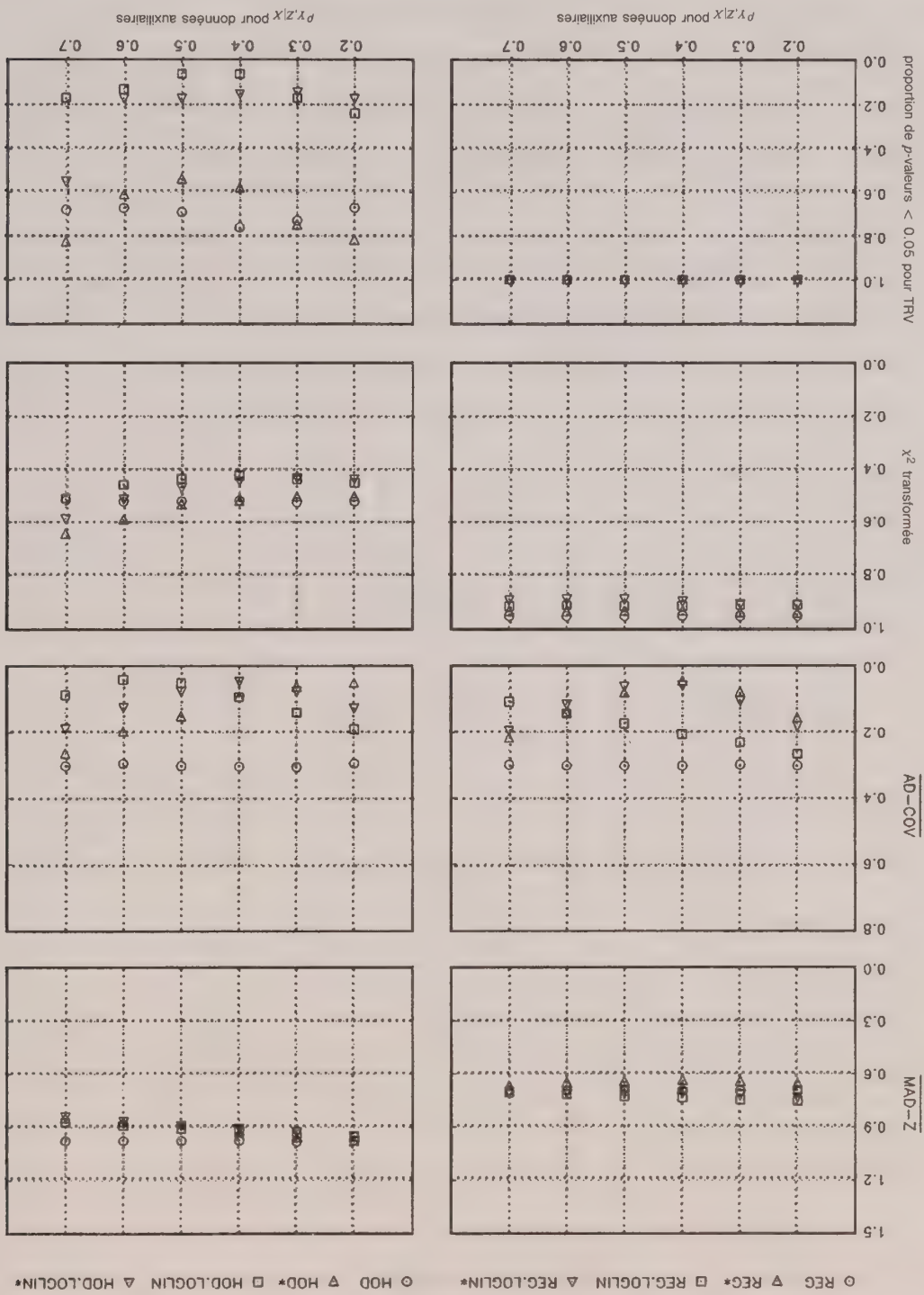


Figure 4. Comparaison de méthodes d'appariement statistique lorsque $p_{Y,Z|X}$ pour le fichier d'information supplémentaire C varie ($p_{Y,Z|X} = .4$ pour les fichiers A et B)

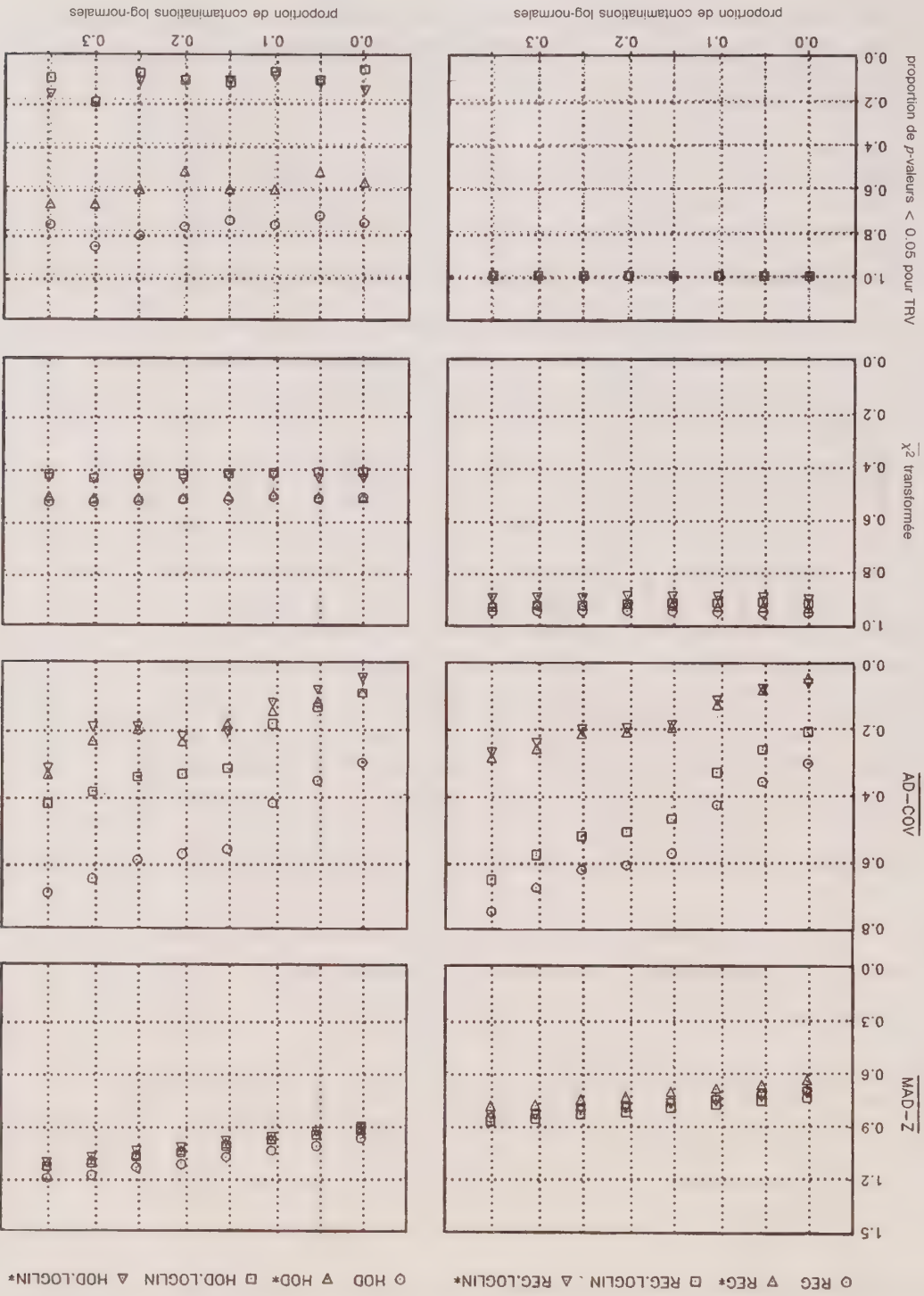


Figure 3. Comparaison de méthodes d'appariement statistique lorsque la proportion de contaminations log-normales varie ($p_{Y,Z|X}$ avant contamination); information supplémentaire non substitutive

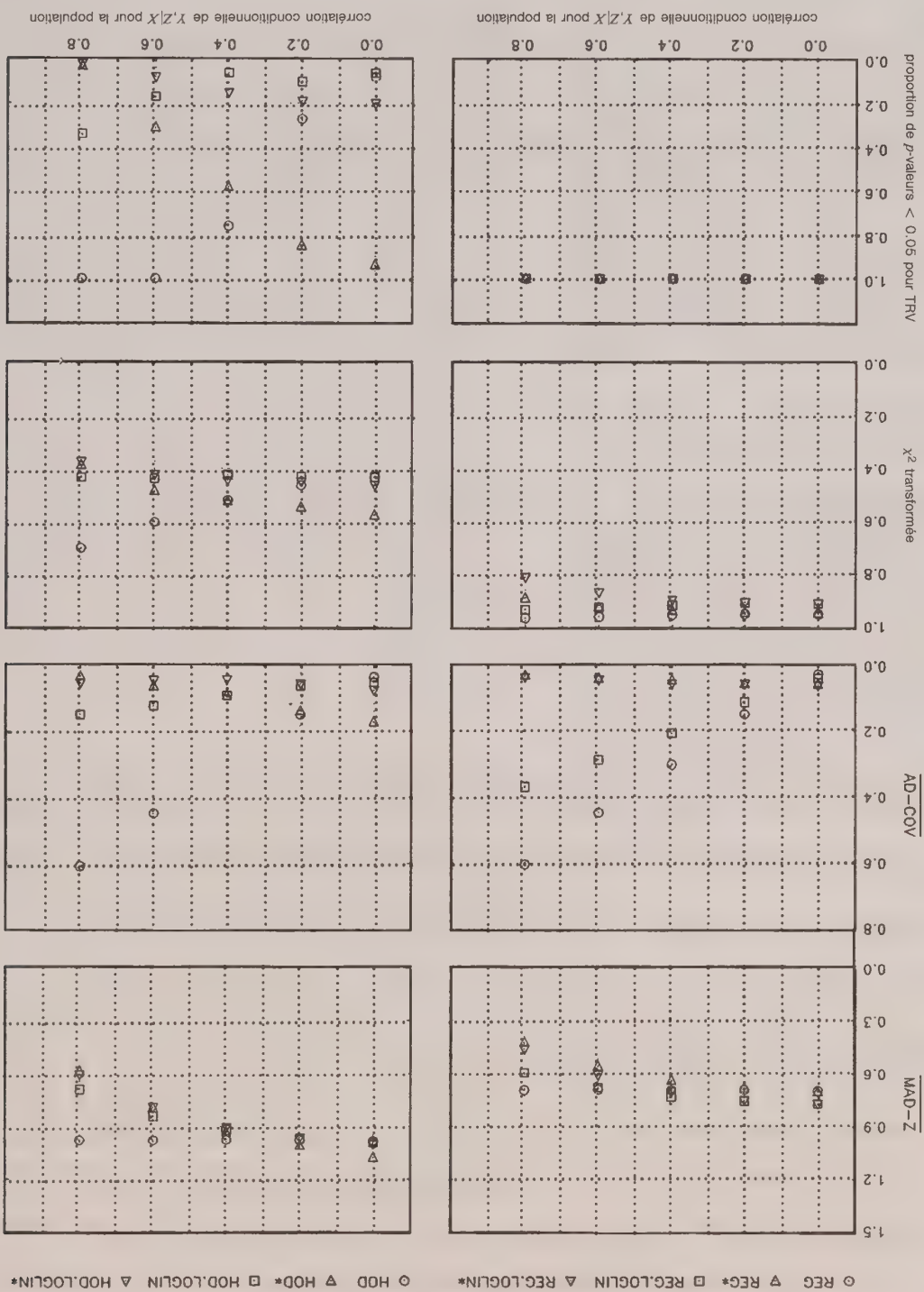


Figure 2. Comparaison de méthodes d'appariement statistique lorsque $\rho_{Y,Z|X}$ varie pour une population symétrique; information supplémentaire non substitutive

Méthodes d'appariement étudiées dans les figures 1 à 5

REG	valeur Z_{int} tirée de la régression de Z par rapport à X , valeur Z_{match} fondée sur la distance (X, Z)
REG*	valeur Z_{int} tirée de la régression de Z par rapport à X et à Y , valeur Z_{match} fondée sur la distance (X, Z)
REG.LOGLIN	valeur Z_{int} tirée de la régression de Z par rapport à X , valeur Z_{match} fondée sur la distance (X, Z) avec des contraintes nominales
REG.LOGLIN*	valeur Z_{int} tirée de la régression de Z par rapport à X et à Y , valeur Z_{match} fondée sur la distance (X, Z) avec des contraintes nominales
HOD	méthode hot-deck utilisant la distance X dans les classes X
HOD*	valeur Z_{int} tirée du fichier C à l'aide de la méthode hot-deck utilisant la distance X , valeur Z_{match} tirée du fichier B à l'aide de la méthode hot-deck utilisant la distance (X, Z)
HOD.LOGLIN	méthode hot-deck utilisant la distance X dans les classes X avec contraintes nominales
HOD.LOGLIN*	valeur Z_{int} obtenue à l'aide de la méthode hot-deck utilisant la distance X avec contraintes nominales, valeur Z_{match} obtenue à l'aide de la méthode hot-deck utilisant la distance (X, Z) dans les classes (X, Y, Z)

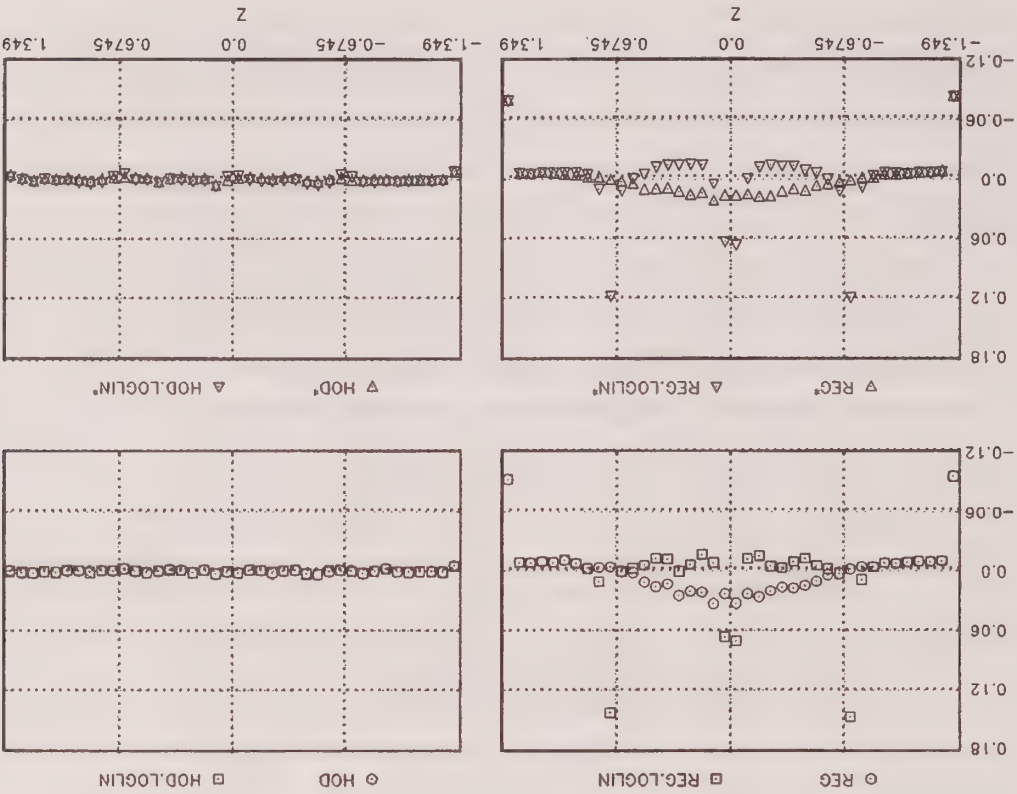


Figure 1. Différence entre les histogrammes marginaux des valeurs Z apparées et des valeurs Z supprimées (données symétriques, $p_{Y,Z|X} = .4$)

Utilisant la même notation que pour la mesure précédente, nous pouvons écrire la formule de la statistique TRV pour la classe $(X^*, Y^*) (i, j)$

$$\begin{aligned} \text{TRV} &= 2 \sum_{k=1}^K \{ (n_{jk} + .5) \ln((n_{jk} + .5) / \\ &\quad (n_{jk} + m_{jk} + 1)) + (m_{jk} + .5) \ln((m_{jk} + .5) / \\ &\quad (n_{jk} + m_{jk} + 1)) \} \\ &\quad + (4n_j + 2K) \ln 2, \end{aligned} \quad (6.5)$$

$$j = 1, \dots, J, \quad k = 1, \dots, K. \quad (6.6)$$

Lorsque les m_{jk} et les n_{jk} proviennent de la même distribution multinomiale, la distribution asymptotique de cette statistique est de type chi carré avec $(K - 1)$ degrés de liberté. On obtient une valeur P globale en faisant la somme des statistiques et des degrés de liberté correspondants pour les classes (X^*, Y^*) qui répondent au critère minimum et en calculant la probabilité que la valeur d'une variable chi carré, étant donné le nombre de degrés de liberté appropriée, soit plus grande que la valeur observée.

7. RÉSULTATS DE L'ÉTUDE DE MONTE CARLO

Dans cette section, nous présentons les résultats de l'étude de simulation. On trouvera une description plus complète de ces résultats dans Singh *et al.* (1990). Les personnes qui voudraient obtenir les tableaux de chiffres qui ont servi à l'élaboration des figures 1 à 5 peuvent en faire la demande. Notre présentation n'insiste pas sur les erreurs types de Monte Carlo des mesures d'évaluation parce qu'elles sont en général relativement faibles; par exemple, les coefficients de variation pour AD-Cov étaient généralement inférieurs à deux pour cent. En outre, les évaluations de différentes méthodes d'appariement devraient normalement être corrélées positivement, de sorte que les erreurs types refléteront beaucoup plus précisément que l'on croit l'indice de la qualité de l'évaluation des diverses méthodes d'appariement par la méthode de Monte Carlo est la régularité générale des tendances observées (voir, par exemple, les figures 2 à 5). Bref, le moindre écart est susceptible de signaler une différence réelle.

7.1 Méthodes n'utilisant pas d'information supplémentaire (REG et HOD)

Les figures 2 à 5 montrent comment l'efficacité des méthodes d'appariement qui appliquent l'HIC est influencée lorsqu'on s'écarte de l'indépendance conditionnelle. Il

semblerait que l'utilisation de méthodes qui appliquent l'HIC crée un biais notable pour la relation conjointe de (X, Y, Z) dans le fichier entiché. On peut constater par exemple dans la figure 2 une détérioration progressive de la situation (sauf pour MAD-Z) lorsque le coefficient vrai de corrélation conditionnelle, $\rho_{Y|Z|X}$, s'éloigne de zéro. Le cas d'exception que représente MAD-Z s'explique probablement par le caractère inconditionnel de cette mesure, qui repose sur une comparaison par unité des valeurs Z appariées et des valeurs Z supprimées, alors que les autres mesures reposent sur des comparaisons des distributions conditionnelles de Z . Il est intéressant de noter dans la figure 2 que lorsque la valeur vraie de $\rho_{Y|Z|X}$ est faible, la méthode HOD*, qui utilise de l'information supplémentaire, peut, en ce qui concerne les mesures d'évaluation de type nominal ou les mesures au niveau global, être moins efficace que la méthode HOD, qui, elle, n'utilise pas cette information. Les conditions dans lesquelles l'utilisation d'informations supplémentaires devient avantageuse dépendraient du degré de précision de ces informations.

7.2 Méthodes utilisant de l'information supplémentaire

Comme il fallait s'y attendre, les résultats de notre étude empirique confirment que l'utilisation d'information supplémentaire est un moyen de compenser la non-vérification de l'HIC. Le degré de compensation dépendrait de la méthode et du genre d'information supplémentaire utilisées. Dans l'introduction, nous avons brossé un tableau du rendement des diverses méthodes étudiées. Maintenant, nous allons examiner plus en détail le rendement de ces méthodes à l'aide des figures 2 à 5.

Parmi les méthodes fondées sur la régression, celles qui utilisent de l'information supplémentaire portant sur la corrélation conditionnelle – notamment REG* et REG, LOGLIN* – ont, pour des populations symétriques, un rendement très acceptable en ce qui concerne les mesures au niveau individuel (c.-à-d. MAD-Z et AD-Cov) (voir figure 2). Elles surclassent aussi les méthodes hot-deck pour des populations asymétriques (figure 3), bien que le biais tende à s'accroître lorsque le degré d'asymétrie augmente. Toutefois, lorsque le coefficient de corrélation conditionnelle pour l'information supplémentaire substitutive varie (figure 4), les méthodes de régression ont un rendement inégal, c.-à-d. qu'elles peuvent être plus efficaces ou moins efficaces que les méthodes hot-deck au niveau individuel. De fait, pour ce qui a trait à la deuxième catégorie d'information supplémentaire substitutive (c'est-à-dire celle avec contamination log-normale; voir figure 5), les méthodes de régression sont généralement un peu moins efficaces que la méthode HOD, LOGLIN en ce qui concerne la mesure AD-Cov). Si nous nous limitons aux méthodes fondées sur la régression, nous pouvons recommander l'usage de la méthode REG* en ce qui regarde les mesures d'évaluation au niveau individuel. Toutefois, pour ce qui a trait aux mesures au niveau global, toutes les méthodes de régression ont un très mauvais rendement. Cela s'explique probablement par l'effet de réduction à la moyenne dont il est question à la sous-section 7.3.

pour l'appariement et la partition à intervalles standard pour les évaluations. La première des quatre mesures d'évaluation repose sur une comparaison par unité des valeurs Z appariées et des valeurs Z supprimées. Cependant, une méthode d'appariement statistique n'a pas pour but de reproduire fidèlement les valeurs Z supprimées mais de produire des valeurs Z qui originent de la même distribution étant donné ce que l'on connaît, en l'occurrence X et Y . Les trois autres mesures d'évaluation reposent plus sur la comparaison des propriétés des distributions conditionnelles de Z .

i) Moyenne des écarts absolus moyens de Z (MAD- Z)

La mesure de rendement la plus simple est l'écart absolu moyen entre les valeurs appariées et les valeurs supprimées de Z pour les enregistrements du fichier A. Nous avons calculé des moyennes de Monte Carlo de ces écarts et les erreurs types correspondantes.

La formule de la statistique MAD- Z pour n importe quelle simulation donnée est

$$\text{MAD-}Z = \sum_{i=1}^I |Z_{s,i} - Z_{m,i}| / 500, \quad (6.1)$$

où $Z_{s,i}$ est la valeur Z supprimée pour le $i^{\text{ème}}$ enregistrement du fichier A, $Z_{m,i}$ est la valeur Z appariée et la sommation est étendue aux 500 enregistrements du fichier A. MAD- Z désigne la moyenne des statistiques MAD- Z pour toutes les simulations.

ii) Moyenne des écarts des covariances (AD-Cov)

La deuxième mesure de rendement est l'écart entre les covariances conditionnelles de Y et Z étant donné X pour le fichier enrichi et le fichier duquel ont été supprimées les valeurs vraies de Z . Nous avons calculé des moyennes de Monte Carlo de ces écarts et les erreurs types correspondantes.

Pour un fichier contenant les variables X , Y et Z , nous pouvons définir la relation

$$\text{Cov}(Y, Z | X) = \text{Cov}(Y, Z) -$$

$$\text{Cov}(X, Y)\text{Cov}(X, Z) / \text{Var}(X), \quad (6.2)$$

où Cov et Var sont les opérateurs de covariance et de variance d'échantillon respectivement. Dans le cas d'une distribution normale multidimensionnelle, cela correspond à la covariance de Y et Z étant donné X ; autrement, on peut le voir comme la covariance des résidus d'une régression linéaire de Y par rapport à X et des résidus d'une régression linéaire de Z par rapport à X . Pour n importe quelle simulation donnée, la statistique AD-Cov serait l'écart entre ces quantités pour le fichier enrichi et

le fichier duquel ont été supprimées les valeurs vraies de Z . Comme pour la première mesure, AD-Cov désigne la moyenne de ces statistiques pour toutes les simulations.

iii) Moyenne des statistiques chi carré (χ^2)

La troisième mesure de rendement, basée sur des comparaisons de type nominal, est une mesure de distance fondée sur le critère chi carré de Pearson. Ce qu'on présente, en fait, est la statistique chi carré moyenne calculée pour les 100 simulations, transformée de manière à être incluse dans l'intervalle (0,1).

La formule de la statistique chi carré est

$$\chi^2 = \sum_{i,j,k} (m_{ijk} - n_{ijk})^2 / (m_{ijk} + .5), \quad (6.3)$$

où m_{ijk} est le nombre d'enregistrements qui appartiennent à la classe X^*i , à la classe Y^*j , et à la classe Z^*k dans le fichier enrichi, n_{ijk} est l'équivalent pour le fichier duquel ont été supprimées les valeurs vraies de Z , et où la sommation est étendue à toutes les classes (X^*, Y^*, Z^*) . Une valeur constante de .5 est ajoutée à tous les dénominateurs de cette somme pour éviter le problème des zéros.

Une fois calculée la moyenne des statistiques chi carré pour 100 simulations, disons χ^2 , cette valeur est transformée de manière à être incluse dans l'intervalle (0,1). La formule de transformation est la suivante (voir Bishop, Fienberg et Holland, 1975, p. 383; le chiffre 500 représente la taille du fichier A)

$$\chi^2 \text{ transformé} = \{ \chi^2 / (\chi^2 + 500) \}^{1/2}. \quad (6.4)$$

iv) Test du rapport des vraisemblances (TRV)

La dernière mesure de rendement est aussi basée sur des comparaisons de type nominal. Pour chaque classe (X^*, Y^*) qui contient un nombre minimum d'observations (dans la présente étude, ce nombre est fixé à 20), on exécute un test du rapport des vraisemblances pour tester l'hypothèse que les valeurs Z de type nominal du fichier enrichi et du fichier amputé proviennent de la même distribution multinomiale. On combine ensuite les tests faits pour les différentes classes (X^*, Y^*) afin d'obtenir une valeur P globale. Ce qui nous intéresse est le nombre de fois, sur 100 simulations, où la valeur P globale était inférieure à .05. Plus cette proportion est élevée, plus il y a de différence entre la distribution des valeurs vraies de Z^* et celle des valeurs appariées étant donné les classes (X^*, Y^*) .

Le critère minimum de 20 unités pour les classes (X^*, Y^*) du fichier A était nécessaire pour que soit acceptable l'approximation chi carré de la distribution de la variable à tester. Si le nombre de classes de Z^* augmentait, il faudrait probablement hausser le critère minimum.

6.1 Plan de l'étude de Monte Carlo

Il faut trois fichiers pour simuler l'appariement statis-

tique: un fichier receveur A, un fichier donneur B et un fichier auxiliaire C. Ces fichiers sont créés synthétiquement à l'aide de distributions spécifiées, chacun d'eux contenant les trois variables X , Y et Z . On supprime la variable Z du fichier A et la variable Y du fichier B. Les valeurs de la variable supprimée du fichier A servent à évaluer le rendement des diverses méthodes d'appariement statisque. Le fichier C peut contenir de l'information sur (X, Z) seulement (si X est supprimée) ou sur les trois variables. Les résultats empiriques présentés dans cet article sont ceux qui correspondent à la situation où C contient de l'information sur (X, Z) seulement, bien que l'autre situation – C contenant de l'information sur (X, Y, Z) – ait été envisagée aussi dans l'étude (voir Singh *et al.* 1990). Des passages de 100 simulations chacun ont été exécutés pour chaque combinaison de paramètres de plan étudiés. On calculait quatre mesures d'évaluation pour chaque simulation, puis on faisait la moyenne de chacune de ces mesures pour les 100 simulations.

Les fichiers A et B étaient toujours tirés de la même distribution, chacun d'eux contenant 500 observations indépendantes et identiquement distribuées. Le fichier C contenait 250 observations qui ne provenaient pas nécessairement de la même distribution que les observations des fichiers A et B; autrement dit, le fichier C pouvait contenir de l'information supplémentaire substitutive ou non substitutive.

Les observations (X, Y, Z) suivaient une distribution normale multidimensionnelle mais dans certains cas, on induisait une contamination log-normale en calculant l'exponentielle de X , de Y et de Z . On déterminait si une observation était contaminée ou non au moyen d'un processus de Bernoulli avec une probabilité déterminée pour tout passage de 100 simulations. Avant la contamination, X , Y et Z étaient des variables normales centrées réduites; de plus, les covariances de (X, Y) et de (X, Z) étaient toujours de .5 tandis que la covariance de (Y, Z) variait d'un passage à l'autre. Par conséquent, la corrélation conditionnelle de Y et Z étant donné X , $\rho_{Y,Z|X}$, variait aussi d'un passage à l'autre.

Pour la plupart des passages, les observations du fichier C suivaient la même distribution que les observations des fichiers A et B. Toutefois, si, dans une application, l'information supplémentaire vient d'une source historique ou se rapporte à des variables de substitution, cette hypothèse peut être déraisonnable. Deux séries de passages ont été exécutées avec de l'information supplémentaire substitutive. Dans la première série, les données supplémentaires avaient un coefficient $\rho_{Y,Z|X}$ différent. Dans la seconde série, les données étaient soumises dans une certaine proportion à une contamination log-normale. Pour les méthodes proposées qui utilisent des contraintes nominales et pour la définition de classes d'appariement pour la méthode HOD, il fallait choisir une partition. Nous en avons utilisé deux. La première, appelée partition à intervalles standard, répartissait les valeurs des variables X , Y et Z dans les classes suivantes: < -1 , $[-1, 0]$, $[0, 1]$, ≥ 1 ;

6.2 Les méthodes d'appariement

autrement dit, la partition était centrée sur la moyenne de la distribution marginale avant contamination, avec des valeurs seuil au centre et à 1 écart-type. La seconde partition, appelée partition à probabilités égales, était semblable à la première sauf que les valeurs seuil coïncidaient avec les quartiles des distributions marginales pré-contamination; autrement dit, la partition consistait dans les classes suivantes: $< -.6745$, $[-.6745, 0]$, $[0, .6745]$, $\geq .6745$. Les partitions étaient définies en fonction des distributions pré-contamination; pour des raisons de simplicité, on utilisait les mêmes partitions lorsque des contaminations log-normales étaient induites. Cependant, il aurait été plus réaliste de faire dépendre les partitions des données.

Nous avons étudié les huit méthodes décrites plus haut. Toutes, sauf REG et HOD, utilisent de l'information supplémentaire. Il existe donc deux versions de chacune de ces méthodes selon que le fichier C renferme de l'information sur (Y, Z) ou (X, Y, Z) . En ce qui concerne les méthodes HOD et HOD.LOGLIN, nous avons considéré trois versions de la hot-deck (notamment, celle fondée sur le rang, la méthode aléatoire et celle fondée sur la distance X) pour tirer des valeurs authentiques de Z du fichier B, mais seuls les résultats de la troisième sont reproduits ici. Pour ce qui a trait aux six autres méthodes, nous avons considéré aussi trois versions de hot-deck (à savoir, distance Z , distance (X, Z) et distance (X, Y, Z)) pour tirer des valeurs authentiques de Z du fichier B, mais nous ne présentons ici, pour des raisons de simplicité, que les résultats relatifs à la distance (X, Z) . La section 7.3 expose sommairement les résultats obtenus avec différentes mesures de distance. On trouvera dans le document de travail de Singh *et al.* (1990) d'autres détails qui ne figurent pas dans le présent article. Notons que si l'on veut utiliser la méthode hot-deck fondée sur la distance (X, Y, Z) pour tirer une valeur authentique de Z du fichier B, il faut tout d'abord déterminer à partir du fichier C des valeurs intermédiaires Y pour B, à l'instar de la valeur Z_{int} pour le fichier A. Notons aussi que nous nous sommes servis de la distance euclidienne à chaque fois qu'une version de la hot-deck fondée sur une mesure de distance était utilisée. Toutefois, nous n'avons pas redressé au préalable les variables par leurs écarts-types respectifs pour des raisons de commodité et parce que toutes les variables de la population fictive avaient des variances communes.

6.3 Les mesures d'évaluation

Quatre mesures ont servi à évaluer le degré d'efficacité des diverses méthodes d'appariement. Toutes les évaluations reposent sur des comparaisons du fichier enrichi avec le fichier duquel ont été supprimées les valeurs vraies de Z . Deux des mesures reposent sur des comparaisons de type nominal, mais il n'est pas nécessaire que les classes qui servent aux évaluations soient celles qu'utilisent les méthodes LOGLIN pour les contraintes nominales. Les résultats présentés ici sont ceux que nous avons obtenus en utilisant la partition à probabilités égales (voir section 6.1)

Tableau 3

Comparaison de huit méthodes d'appariement permettant de compléter le fichier A

Fichier A		Versions de la méthode REG								Versions de la méthode HOD			
		Valeurs Z apparées											
		X	Y	Z	M1	M2	M3	M4	M5	M6	M7	M8	
Mesures d'évaluation		0.90	1.19	-1.01	0.79	0.81	0.76	0.78	0.79	0.85	0.78	0.78	
MAD-Z		0.50	1.24	1.34	-0.49	0.85	-0.42	-0.38	-0.49	0.85	-0.42	0.85	
χ^2		1.63	1.03	1.53	0.31	0.99	0.31	0.99	0.31	0.99	0.31	0.99	
		1.38	0.79	0.32	0.31	0.99	0.31	0.99	0.99	-0.42	0.99	-0.42	
		1.32	0.61	2.08	0.31	0.99	0.31	0.99	1.24	-0.42	1.24	-0.42	
		1.90	-1.07	0.01	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	
		0.95	-2.15	-1.26	-0.42	-1.16	-0.42	-1.16	-0.42	-1.16	-0.42	-1.16	
		0.06	-1.29	0.33	-0.38	-0.38	-0.36	-0.38	-0.38	0.85	-0.49	0.36	
		0.37	-0.04	-1.18	-0.38	-0.38	-0.38	0.36	-0.49	-0.49	-0.83	-0.83	
		-0.56	0.00	1.06	-0.83	-0.83	-0.83	0.36	-0.69	0.36	-0.69	0.36	
		-0.81	0.56	-0.76	-0.69	0.36	-0.69	0.36	-0.38	0.36	-0.38	0.36	
		-0.42	0.62	-0.44	-0.38	0.36	0.36	-0.38	0.36	-0.38	0.36	-0.38	
		-0.09	-0.26	0.19	-0.38	-0.38	-0.69	0.36	-0.83	-0.69	-0.83	-0.69	
		-0.77	-0.33	0.16	-0.69	-0.69	-0.69	-0.69	-0.69	-0.69	-0.69	-0.69	
		-0.86	-0.32	-0.97	-0.69	-0.69	-0.69	-0.69	-0.69	-0.69	-0.69	-0.69	

Tableau 4

Étapes de la méthode M8 (HOD.LOGLIN*)

Case (X,Y)	X	Y	Ordre d'appariement	Dist. Y	Z _{int}	Dist. (X,Z)	Z _{match}
X < 0	-0.86	-0.32	(A1)	2	0.08	(C1)	0.12
Y < 0	-0.77	-0.33	(A2)	1	0.07	(C1)	0.19
X < 0	-0.42	0.62	(A4)	2	0.05	(C4)	0.56
Y > 0	-0.81	0.56	(A5)	1	0.01	(C8)	0.50
X > 0	0.37	-0.04	(A7)	4	0.36	(C1)	0.14
Y < 0	0.06	-1.29	(A8)	1	0.03	(C6)	0.57
X > 0	0.95	-2.15	(A9)	2	0.18	(C2)	1.66
Y > 0	1.90	-1.07	(A10)	3	0.25	(C6)	0.40
X > 0	1.32	0.61	(A11)	1	0.06	(C4)	0.37
Y > 0	1.38	0.79	(A12)	3	0.12	(C4)	0.43
X > 0	1.63	1.03	(A13)	5	0.28	(C9)	0.25
Y > 0	0.50	1.24	(A14)	2	0.07	(C9)	0.41
X > 0	0.90	1.19	(A15)	4	0.12	(C9)	0.29

Nota : L'ordre d'appariement des enregistrements des fichiers A et C est renouvelé pour chaque case (X,Y) selon les contraintes nominales établies dans le tableau 2(d).

Tableau 2

Distributions de variables nominales pour les fichiers A, B et C suivant la partition donnée $2 \times 2 \times 2$

(a)	Fichier A			Fichier B			Fichier C		
	$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$	
	3	4	5	1	4	4	3	1	3
	$Y < 0$		$Y \geq 0$	$Z < 0$		$Z \geq 0$	$Z < 0$		$Z \geq 0$

(b)	Tableau non redressé du fichier A			Tableau redressé du fichier B			Tableau redressé du fichier C		
	$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$	
	3	4	5	1.5	4.5	4.5	1.85	3.85	4.15
	$Y < 0$		$Y \geq 0$	$Z < 0$		$Z \geq 0$	$Z < 0$		$Z \geq 0$

(c)	Tableau obtenu à partir d'un tableau $2 \times 2 \times 2$ de uns "balayé par itération" en vue d'une concordance avec les fréquences marginales du tableau 2(b)								
	$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$	
	2.55	2.60	1.95	0.45	1.40	3.10	1.05		
	$Y < 0$		$Y \geq 0$	$Z < 0$		$Z \geq 0$	$Z < 0$		$Z \geq 0$

(d)	Contraintes nominales établies par l'arrondissement aléatoire des éléments du tableau 2(c)								
	$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$		$X < 0$	$X \geq 0$	
	2	2	2	1	2	3	1		
	$Y < 0$		$Y \geq 0$	$Z < 0$		$Z \geq 0$	$Z < 0$		$Z \geq 0$

6. ANALYSE EMPIRIQUE DES MÉTHODES D'APPARIEMENT PROPOSÉES

Dans cette section, nous décrivons une analyse empirique qui prend la forme d'une étude de simulation approfondie qui utilise des données fictives tirées de distributions multidimensionnelles symétriques et asymétriques. La symétrie provient de distributions normales tandis que la dissymétrie est introduite par des contaminations engendrées par des distributions log-normales multidimensionnelles. Nous utilisons des données fictives afin d'exercer un contrôle sur tous les paramètres pertinents, y compris ceux qui définissent les relations conjointes des différentes variables. On peut alors évaluer les diverses méthodes d'appariement lorsque les relations conjointes s'écartent d'une manière systématique de l'indépendance conditionnelle. On peut aussi comparer ces méthodes lorsque la distribution d'où sont tirées les données s'écarte de la symétrie. Nous avons obtenu de l'information supplémentaire substitutive en modifiant les paramètres de la distribution normale d'où était tiré le fichier C ou en induisant des contaminations log-normales. Nous avons donc quatre types de problème d'appariement: d'une part, deux distributions, l'une symétrique et l'autre asymétrique, avec des données du fichier C qui sont non substitutives et d'autre part, deux distributions symétriques avec, chacune, une catégorie de données substitutives du fichier C différente. La programmation s'est faite sur micro-ordinateur à l'aide du logiciel GAUSS.

Tableau 1

Données des fichiers A, B, C

Identificateur d'enregistrement	Fichier A X	Y	Identificateur d'enregistrement	Fichier B X	Y	Identificateur d'enregistrement	Fichier C Y	Z
A1	-0.86	-0.32	B1	-0.95	-0.69	C1	-0.40	-0.60
A2	-0.77	-0.33	B2	-0.64	-0.83	C2	-2.33	-2.81
A3	-0.09	-0.26	B3	-1.58	-0.11	C3	-0.79	-0.47
A4	-0.42	0.62	B4	-0.42	0.36	C4	0.67	-0.29
A5	-0.81	0.56	B5	0.97	-0.42	C5	-0.65	1.19
A6	-0.56	0.00	B6	1.09	-1.16	C6	-1.32	0.05
A7	0.37	-0.04	B7	0.44	-0.49	C7	-0.55	0.70
A8	0.06	-1.29	B8	0.14	-0.38	C8	0.55	0.66
A9	0.95	-2.15	B9	1.33	1.24	C9	1.31	1.12
A10	1.90	-1.07	B10	0.80	0.85	C10	1.46	2.58
A11	1.32	0.61	B11	1.60	0.31			
A12	1.38	0.79	B12	1.42	0.99			
A13	1.63	1.03						
A14	0.50	1.24						
A15	0.90	1.19						

les tableaux redressés des fichiers B et C; ce redressement est effectué de sorte qu'il y ait concordance entre ces tableaux et les fréquences marginales pertinentes, comme nous l'expliquons dans la section 2. Le tableau 2(c) correspond au tableau tridimensionnel obtenu à la suite d'un balayage par itération et le tableau 2(d) montre les contraintes nominales obtenues à la suite de l'arrondissement aléatoire des éléments du tableau 2(c), comme nous l'avons expliqué dans la section 2.

Les huit méthodes ont été appliquées aux données du tableau 1 et les résultats de l'appariement figurent dans le tableau 3 avec les valeurs vraies de Z, qui avaient été supprimées dans le tableau 1.

Les mesures d'évaluation indiquées dans le tableau 3 ont été mentionnées brièvement dans l'introduction et sont expliquées en détail dans la section 6. La partition utilisée pour la mesure χ^2 est la même que celle qui a servi à la détermination des contraintes nominales. Comme la partition n'a pas été modifiée pour l'évaluation, il convient de noter que les valeurs de χ^2 pour M3, M4, M7 et M8 seront identiques. Notons aussi que les mesures d'évaluation ne sont présentées ici que dans le but d'illustrer le calcul du rendement relatif des méthodes puisqu'elles ne reposent que sur un petit échantillon.

La méthode M8 (HOD.LOGLIN*) est celle qui comporte le plus grand nombre de calculs; la table 4 en décrit les étapes. À partir de ce tableau, on peut se représenter assez facilement les étapes requises pour les autres méthodes.

Dans la méthode de Paass, on commence par déterminer la distribution conditionnelle de Z pour chaque couple (X, Y) en A en la représentant au moyen d'un ensemble de K valeurs Z obtenu par une régression non paramétrique. Autrement dit, on ajoute K valeurs Z à chaque couple (X, Y) . En deuxième lieu, on détermine, pour chaque couple (X, Y) en A , une valeur Z authentique - Z_{match} - tirée du fichier B , qui est considérée comme la plus proche suivant la distance (X, Z) . On obtient ainsi le triplet (X, Y, Z_{match}) pour le fichier qui a fait l'objet de l'appariement (fichier A). Si le fichier C contient de l'information sur (Y, Z) , on détermine les distributions conditionnelles pour le fichier A à l'aide du processus itératif suivant: choisir K valeurs initiales comme "plus proches voisins" pour Z , dans le fichier A , Y , dans le fichier B , et X , dans le fichier C ; cela peut se faire par la méthode d'imputation hot-deck ordinaire. Or, chaque cycle consiste à établir des distributions conditionnelles pour les éléments (X, Y) en A à l'aide de l'information contenue dans le fichier C , plus précisément en mettant à jour de façon conforme K valeurs Z du fichier A à l'aide de valeurs Z du fichier C en utilisant la distance (X, Y) , puis en mettant à jour K valeurs Y du fichier B à l'aide de valeurs Y du fichier A en utilisant la distance (X, Z) et enfin, en mettant à jour K valeurs X du fichier C à l'aide de valeurs X du fichier B en utilisant la distance (Y, Z) . Ce cycle est répété jusqu'à ce que la différence maximale entre des statistiques de la distribution tridimensionnelle de (X, Y, Z) obtenue par itérations successives (par ex.: matrice de covariance) soit inférieure à une limite donnée. À la convergence, chaque fichier compte K nouvelles valeurs qui représentent les distributions conditionnelles respectives. Par ailleurs, si le fichier C contient de l'information sur (X, Y, Z) , le processus est non itératif. Il suffit de se servir du fichier C pour obtenir K valeurs Z de A en utilisant la distance (X, Y) , puis de déterminer une valeur Z_{match} , tirée du fichier B , pour chaque couple (X, Y) en A en utilisant la distance (X, Z) . Toutefois, Paass n'envisage pas cette possibilité dans son article.

Dans l'étude empirique que nous exposons ici, nous n'avons pas utilisé la version itérative de la méthode de Paass décrite ci-dessus lorsque le fichier C contenait de l'information sur (Y, Z) parce qu'elle comporte de nombreux calculs. Nous avons plutôt utilisé une version non itérative simplifiée où $K = 1$. Cette méthode, désignée par HOD^* , consiste en deux étapes.

HOD^* (étape 1) Pour chaque couple (X, Y) en A , déterminer par la méthode hot-deck une valeur intermédiaire Z_{int} à partir du fichier C en utilisant la distance Y si l'information supplémentaire porte sur (Y, Z) et la distance euclidienne (X, Y) si l'information supplémentaire porte sur (X, Y, Z) .

HOD^* (étape 2) Remplacer chaque triplet (X, Y, Z_{match}) obtenu à l'étape 1 par (X, Y, Z_{match}) , où Z_{match} est tiré du fichier B à l'aide de la méthode hot-deck avec la distance euclidienne (X, Z) .

S'il n'existe pas de fichier C , on peut se servir de la méthode HOD , appliquée avec l'HIC. Cette méthode comporte deux étapes:

HOD (étape 1) Déterminer des classes X appropriées, comme dans l'imputation hot-deck ordinaire.

HOD (étape 2) Pour chaque couple (X, Y) en A , imputer une valeur Z observée, tirée de la classe X correspondante dans le fichier B , en utilisant la méthode hot-deck avec la distance X .

4. MODIFICATION DES MÉTHODES D'APPARIEMENT PAR L'INTRODUCTION DE CONTRAINTES NOMINALES

Nous proposons de modifier les méthodes d'appariement REG, REG*, HOD et HOD^* en définissant des contraintes de type nominal pour les valeurs Z que l'on tire du fichier B pour compléter le fichier A . Ces contraintes ont pour but de conserver les liens catégoriques (définis par des modèles log-linéaires) suivant une partition de (X, Y, Z) acceptable dans le cas du fichier enrichi. On détermine ces liens en combinant des informations des fichiers A, B et C . La notion d'appariement sous contrainte nominale repose sur la méthode d'imputation log-linéaire (voir Singh 1988; Singh *et al.* 1988). Dans ce cas, les contraintes peuvent dépendre de l'information supplémentaire qui peut servir à estimer la distribution conditionnelle de variables nominales, ou certains aspects de cette distribution, mais qui n'a pas le niveau de qualité voulu pour estimer la distribution conditionnelle intégrale.

Nous débutons avec une partition adéquate des variables X, Y et Z . Posons X^*, Y^*, Z^* comme les variables nominales correspondantes (après transformation). On peut alors paramétrer la distribution des proportions par case pour le tableau (X^*, Y^*, Z^*) au moyen d'un modèle log-linéaire

$$\log p_{ijk} = u + n_i + n_j + n_k$$

$$(4.1) \quad + n_{12ij} + n_{13ik} + n_{23jk} + n_{123ijk},$$

où p_{ijk} est la proportion relative à la case (i, j, k) et les nombres initiaux 1, 2 et 3 désignent respectivement X^*, Y^* , et Z^* . Il convient de souligner que l'équation de paramétrage (4.1) vaut pour des distributions arbitraires des variables originales (X, Y, Z) . Evidemment, les fichiers A et B ne contiennent pas d'information sur les effets à deux facteurs n_{23} et les effets à trois facteurs n_{123} . Si ces effets sont posés égal à zéro, cela revient à poser l'hypothèse de l'indépendance conditionnelle au sens qualitatif, c.-à-d. $Y^* \perp Z^* \mid X^*$. Cependant, si nous disposons d'informations supplémentaires dans le fichier C , nous pouvons nous passer de cette hypothèse parce qu'il est possible d'estimer les paramètres n_{23} et n_{123} à l'aide du fichier C . Par conséquent, peu importe la forme de la distribution conjointe de (X, Y, Z) , le modèle log-linéaire ci-dessus offre une façon homogène d'évaluer, à tout le moins au sens qualitatif, la distanciation par rapport à l'HIC. Par ailleurs, dans la régression linéaire, la distanciation par rapport à

moymenne μ_Z , et de l'écart-type σ_Z du fichier B et des estimations du coefficient de corrélation $\rho_{X,Y}$, des moyennes μ_X, μ_Y et des écarts-types σ_X, σ_Y du fichier A. Par conséquent, on ne se servira du fichier B que si le fichier A ne contient pas tous les renseignements voulus sur la variable étudiée et on ne se servira du fichier C que si les fichiers A et B ne contiennent pas toute l'information voulue. Nous supposons donc l'existence d'une échelle de fiabilité ou de pertinence pour les fichiers A, B et C. Une telle échelle n'existait pas dans l'étude de Rubin. Nous pouvons alors calculer les estimations voulues à l'aide des équations

$$\beta_2 = \rho_{Y,Z|X} \frac{\sigma_{Y|X}}{\sigma_Z}, \quad \beta_1 = \rho_{X,Z|Y} \frac{\sigma_{X|Y}}{\sigma_Z}, \quad (3.1)$$

$$\beta_0 = \mu_Z - \beta_1 \mu_X - \beta_2 \mu_Y, \quad (3.2)$$

où

$$\sigma_{Z|X} = (1 - \rho_{X,Z}^2)^{1/2} \sigma_Z, \quad \sigma_{Y|X} = (1 - \rho_{X,Y}^2)^{1/2} \sigma_Y,$$

$$\sigma_{Z|Y} = (1 - \rho_{Y,Z}^2)^{1/2} \sigma_Z, \quad \sigma_{X|Y} = (1 - \rho_{Y,X}^2)^{1/2} \sigma_X, \quad (3.3)$$

et où $\rho_{X,Z|Y}$ est déterminé à l'aide de la formule habituelle après qu'on a calculé $\rho_{Y,Z}$ au moyen de la formule (2.1), c.-à-d.

$$\rho_{X,Z|Y} = (\rho_{X,Z} - \rho_{X,Y} \rho_{Y,Z}) (1 - \rho_{X,Y}^2)^{-1/2}. \quad (3.4)$$

Notons que suivant l'hypothèse de la normalité, les écarts par rapport à l'indépendance conditionnelle sont représentés par le paramètre $\rho_{Y,Z|X}$. S'il y a indépendance conditionnelle, $\rho_{Y,Z|X} = 0$ et le modèle (3.1) se réduit à la régression linéaire simple de Z par rapport à X, c.-à-d.

$$E(Z|X) = \beta_0 + \beta_1 X, \quad V(Z|X) = \sigma^2, \quad (3.5)$$

ce modèle peut être spécifié par la combinaison de données des fichiers A et B ou du fichier B seulement. Les formules (3.2) se ramènent à

$$\beta_2 = 0, \quad \beta_1 = \rho_{X,Z}, \quad \beta_0 = \mu_Z - \beta_1 \mu_X. \quad (3.6)$$

Lorsque le fichier C contient de l'information sur $\rho_{Y,Z}$ seulement, on peut estimer facilement les paramètres de (3.1) d'une manière comparable à celle ci-dessus en combinant des données des fichiers A, B et C.

Une fois qu'on a défini le modèle de régression, on peut appliquer la méthode REG* en deux étapes. La seconde étape est importante puisque nous voulons avoir des valeurs authentiques de Z de sorte que les relations entre les éléments de la variable multidimensionnelle Z soient préservées.

REG* (étape 1) Pour chaque couple (X, Y) en A, déterminer une valeur intermédiaire Z_{int} à l'aide du modèle de régression (3.1).

REG* (étape 2) Remplacer chaque triplet (X, Y, Z_{int}) obtenu à l'étape 1 par (X, Y, Z_{match}) , où Z_{match} désigne une valeur authentique de Z tirée de B qui est le plus près

possible de la valeur intermédiaire selon la distance euclidienne en (X, Z) , où les éléments X et Z sont pondérés par leurs écarts-types respectifs. Autrement dit, on se sert de la méthode hot-deck pour déterminer les valeurs authentiques (observées). C'est ce que Rubin appelle l'"appariement par régression avec moyenne prédictive"; voir Little et Rubin (1987).

Ce qui distingue aussi la méthode décrite ci-dessus de celle présentée dans Rubin (1986) est que, selon celle-ci, on prédit une valeur Y pour les enregistrements du fichier B au moyen d'équations analogues à celles de (3.1), puis on prédit les valeurs Z correspondantes; ensuite, on apparie les enregistrements du fichier A à ceux du fichier B en fonction de la différence entre les valeurs Z prédites. S'il n'y a pas d'information supplémentaire, on peut se servir de la méthode REG, appliquée avec l'HIC. Cette méthode comporte aussi deux étapes:

REG (étape 2) Même chose que pour REG*.

Là aussi, on note des différences avec la méthode décrite dans Rubin (1986); selon celle-ci, on prédit une valeur Z pour les enregistrements du fichier B au moyen des équations (3.5), puis on apparie les enregistrements du fichier A à ceux du fichier B en fonction de la différence entre les valeurs Z prédites. Dans le cas qui nous occupe, où X est une variable unidimensionnelle, cela équivaut à un appariement par rapport à X.

3.2 Méthode hot-deck

Nous décrivons tout d'abord une méthode hot-deck qui utilise de l'information supplémentaire. Il s'agit d'une version de la méthode proposée par Paas (1986). Nous parlons ici de régression non paramétrique. Dans la régression paramétrique, la distribution conditionnelle de Z étant donnée X et Y est définie au sens large par des fonctions de moyenne et de variance qui font intervenir un petit nombre de paramètres. Dans la régression non paramétrique, on se sert des techniques d'estimation non paramétrique de la densité pour estimer la distribution conditionnelle de Z. Par exemple, en ce qui a trait à la méthode d'estimation de la densité dite du plus proche voisin, on détermine, pour chaque couple (X, Y) , K "plus proches voisins" (par rapport à une fonction de distance comme la distance euclidienne en (X, Y) ; cet échantillon (probablement pondéré) de K "voisins", où K est un nombre entier spécifié convenablement, sert à représenter la distribution conditionnelle. Par conséquent, $P(Z \in U | X, Y)$ peut être spécifiée comme une espérance conditionnelle,

$$E(I_U(Z) | X, Y) = \sum_{i=1}^K w_i(X, Y) I_U(Z_i), \quad (3.7)$$

où les w_i sont des poids qui diminuent à mesure que s'accroît la distance entre (X_i, Y_i) et (X, Y) , et I_U est la fonction indicatrice pour l'ensemble U.

Deuxièmement, l'information supplémentaire n'a pas besoin d'être parfaite, c'est-à-dire qu'elle peut être définie dans une certaine mesure. Par exemple, elle peut venir d'une source périmée (une enquête ou un recensement antérieurs peut-être) – mais les renseignements qui nous intéressent dans cette source peuvent être, eux, tout à fait pertinents – ou elle peut représenter à tout le moins une amélioration par rapport à l'HIC, qui s'imposerait de toute manière. Par ailleurs, l'information supplémentaire peut avoir trait à un ensemble de variables substitutives qui sont supposées évoluer de la même manière que les variables étudiées.

Il peut y avoir de l'information supplémentaire de macro-niveau ou de micro-niveau. L'information de macro-niveau peut prendre la forme de coefficients de corrélation ou de proportions par case (de type nominal) ou encore d'autres paramètres. Si l'information supplémentaire contenue dans le fichier C porte sur la corrélation conditionnelle de Y et Z étant donné X, c'est-à-dire $\rho_{Y,Z|X}$, elle peut être combinée avec l'information des fichiers A et B concernant les corrélations de X et Y et de X et Z pour estimer la corrélation inconditionnelle de Y et Z au moyen de l'équation

$$\rho_{Y,Z} = \rho_{X,Y} \rho_{X,Z} + \rho_{Y,Z|X} (1 - \rho_{X,Y}^2)^{1/2} (1 - \rho_{X,Z}^2)^{1/2}. \quad (2.1)$$

Or, on peut se servir des données des fichiers A et B pour faire une régression linéaire de Z par rapport à X et à Y en ce qui concerne la méthode REG* (voir section 3.1). Si l'information supplémentaire dont on dispose ne porte que sur la corrélation inconditionnelle de Y et Z, elle peut néanmoins servir de la même façon.

L'information supplémentaire de macro-niveau du fichier C peut aussi prendre la forme d'une distribution de variable nominale, comme la distribution de (X^*, Y^*, Z^*) , où * désigne la transformation de la variable initiale en une variable nominale. S'il s'agissait de variables nominales dès le départ, on n'aurait pas à les transformer. On obtient le tableau de fréquences nécessaire pour les méthodes d'appariement sous contrainte nominale en "balayant par itération" le tableau (X^*, Y^*, Z^*) qui correspond au fichier C de manière que les tableaux marginaux (X^*, Y^*) et (X^*, Z^*) concordent respectivement avec le tableau (X^*, Y^*) du fichier A et le tableau (X^*, Z^*) du fichier B. Notons que ce dernier tableau devra avoir été "balayé par itération" au préalable de manière que ses fréquences marginales X^* concordent avec celles du fichier A. Dans la détermination des contraintes nominales, la méthode du "balayage par itération" préserve les liens (Y^*, Z^*) et (X^*, Y^*, Z^*) du tableau (X^*, Y^*, Z^*) du fichier C. L'opération d'ajustement que nous avons effectuée plus haut sur ce tableau est acceptable parce que les données sur la distribution (X^*, Y^*) du fichier A et la distribution (X^*, Z^*) du fichier B sont réputées plus exactes ou plus pertinentes que celles du fichier C. Si l'on ne connaît (ou n'utilise) que la distribution (Y^*, Z^*) du fichier C, il est possible de modifier la procédure de balayage par

La section suivante traite de l'utilisation de l'information supplémentaire dans l'appariement statistique. Elle décrit aussi l'utilisation de l'information supplémentaire de micro-niveau. Lorsque ce genre d'information existe, on peut, la plupart du temps, ramener cette information au macro-niveau par des totalisations et obtenir des données fiables sur les coefficients de corrélation et les proportions par case (de type nominal). Le succès de cette opération dépendra en partie de la taille du fichier de micro-données.

3. APPARIEMENT STATISTIQUE

3.1 Méthode de régression

Nous allons décrire tout d'abord une méthode de régression qui utilise de l'information supplémentaire. Il s'agit d'une version de la méthode proposée par Rubin (1986). On suppose une forme paramétrique pour la régression de Z par rapport à X et à Y et on estime ensuite les paramètres correspondants à l'aide des données des fichiers A, B et C. Par exemple, dans le cas d'une régression linéaire, nous avons le modèle

$$E(Z|X, Y) = \beta_0 + \beta_1 X + \beta_2 Y, \\ V(Z|X, Y) = \sigma^2, \quad (3.1)$$

où β_0, β_1 , et β_2 sont estimés à l'aide d'équations semblables aux équations des moindres carrés habituelles et par une combinaison adéquate de données des fichiers A, B et C. Nous décrivons ci-dessous une méthode pour estimer ces paramètres, qui diffère quelque peu de celle présentée dans Rubin (1986). Si le fichier C contient de l'information sur $(X, Y$ et $Z)$, on peut obtenir des estimations du coefficient de corrélation conditionnelle $\rho_{Y,Z|X}$ du fichier C, des estimations du coefficient de corrélation $\rho_{X,Z}$, de la

principalement Ruggles, Ruggles et Wolff (1977), Paass et Wauschkuhn (1980), Bart, Stewart et Turner (1981) de même que Rodgers et DeVoi (1982). Dans Paass (1986), on fait une excellente analyse des tests empiriques portant sur la qualité des méthodes d'appariement.

Toutes les études mentionnées ci-dessus confirment que l'HIC peut représenter une contrainte sérieuse. Elles font ressortir la nécessité d'introduire de nouvelles informations dans le processus d'appariement. Peu d'études empiriques considèrent l'utilisation d'informations supplémentaires et l'incidence de l'HIC; Paass (1986) fait une évaluation en se servant uniquement de données fictives, tandis que Armstrong (1989) effectue des simulations à l'aide de données fictives et de données réelles. Nous pouvons considérer que la présente étude s'inscrit dans la suite de ces études en ce sens que de nouvelles méthodes y sont présentées et que l'éventail des distributions de population correspondantes est assez large.

Le plan de cet article est le suivant. La section 2 présente différents types d'information supplémentaire. La section 3 contient une analyse succincte de diverses méthodes d'appariement qui utilisent de l'information supplémentaire et la section 4 présente les modifications proposées sous forme de contraintes nominales. La section 5 sert à illustrer les diverses méthodes d'appariement par un exemple numérique simple. La section 6 contient la description du plan de l'étude empirique portant sur les méthodes d'appariement proposées et la section 7 contient l'analyse des résultats. Enfin, la section 8 renferme les conclusions de l'étude et propose des voies de recherche.

2. TYPES D'INFORMATION SUPPLÉMENTAIRE

Bien qu'il n'existe pas de fichier de micro-données courant et suffisamment gros qui contienne de l'information sur tout l'ensemble des variables, il peut exister une source auxiliaire qui renferme de l'information sur certaines des relations conjointes entre les éléments de l'ensemble de variables (X, Y, Z) ou les éléments du sous-ensemble (X, Z). Si tel est le cas, on peut intégrer cette source au processus d'appariement pour éviter de poser l'HIC et pour améliorer la qualité du fichier pour compléter par une réduction les variables du fichier qui a fait l'objet de l'appariement. Cette information supplémentaire peut venir de diverses sources et se présenter sous plusieurs formes. Comme cette information a pour seul but de nous permettre de ne pas poser l'HIC, nous en limitons l'usage en ce sens qu'elle ne peut jamais remplacer l'information contenue dans les fichiers receveur ou donneur. Autrement dit, il s'agit de tirer de la source auxiliaire des informations que l'on ne retrouve pas dans les fichiers de base. L'opération se fait en toute conformité avec les règles de protection du secret statistique qui s'appliquent à la source auxiliaire et suppose que cette source peut être le produit d'une petite enquête menée à des fins particulières ou un fichier de données confidentiel.

lorsqu'on s'écarte quelque peu de l'hypothèse d'indépendance conditionnelle, l'utilisation d'information supplémentaire n'accroît pas nécessairement l'efficacité de la méthode HOD en ce qui concerne les mesures d'évaluation au niveau global. Cette constatation devrait avoir une incidence notable dans les cas où il est difficile d'obtenir de l'information supplémentaire.

ii) La méthode REG* offre un rendement très acceptable en ce qui concerne les mesures d'évaluation au niveau individuel. On observe tout le contraire en ce qui a trait aux mesures au niveau global; cela est probablement attribuable au phénomène de "réduction à la moyenne", propre aux techniques de régression.

iii) La méthode HOD* est de beaucoup supérieure à la méthode REG* au niveau global mais elle est, en général, un peu moins efficace au niveau individuel.

iv) De façon générale, les contraintes nominales accroissent l'efficacité des méthodes REG* et HOD*. Plus précisément, en ce qui a trait aux mesures d'évaluation au niveau global, la méthode REG.LOGLIN* devient un peu plus efficace grâce à ces contraintes tandis que la méthode HOD.LOGLIN* devient beaucoup plus efficace. En ce qui concerne les mesures au niveau individuel, l'efficacité des deux méthodes demeure essentiellement la même.

v) En ce qui a trait aux mesures au niveau global, la méthode HOD.LOGLIN basée uniquement sur de l'information supplémentaire de type nominal est généralement plus efficace que la méthode HOD.LOGLIN* basée sur de l'information supplémentaire de micro-niveau. En revanche, pour ce qui a trait aux mesures au niveau individuel, la méthode HOD.LOGLIN est relativement moins efficace que la méthode HOD.LOGLIN*. Cette constatation peut être importante du point de vue pratique car la méthode HOD.LOGLIN est beaucoup moins exigeante sur le plan du calcul que la méthode HOD.LOGLIN* et elle ne nécessite pas d'information supplémentaire de micro-niveau. La méthode REG.LOGLIN n'a pas une efficacité comparable, à cause probablement du phénomène de "réduction à la moyenne".

vi) S'il s'agit d'informations supplémentaires substitutives ou complémentaires, leur utilisation peut néanmoins être profitable. Dans ces circonstances, la méthode HOD.LOGLIN donne des résultats très acceptables et, de fait, est passablement robuste à l'égard d'une information supplémentaire déficiente. Notons que comme cette méthode n'utilise que de l'information touchant les liens catégoriques entre les données supplémentaires, il semble raisonnable de penser qu'elle sera peu influencée par le fait que les données du fichier C sont substitutives ou quelque peu complémentaires. On ne peut en dire autant, toutefois, de la méthode REG.LOGLIN. Il convient de souligner que plusieurs études empiriques ont été faites dans le passé afin d'évaluer les méthodes d'appariement statistique. Parmi celles qui ne considèrent pas l'utilisation d'informations supplémentaires, notons

l'égard d'informations supplémentaires de qualité inférieure ou imparfaites dans le fichier C. Si l'information supplémentaire est sous forme de distribution de variable nominale et qu'elle n'est pas de l'information de micro-niveau, on peut modifier les méthodes d'appariement fondées sur l'HIC en introduisant des contraintes nominales; dans ce cas, l'HIC est utilisée uniquement pour les classes de valeurs (X et Y). Par exemple, les méthodes d'imputation ordinaires REG et HOD, qui peuvent servir à l'appariement sans que l'on tienne compte de Y , peuvent aboutir à des versions modifiées REG.LOGLIN et HOD.LOGLIN. Ces deux dernières sont aussi considérées dans cet article.

Il convient de souligner que les méthodes d'appariement assujetties à des contraintes nominales diffèrent des méthodes d'appariement sous contrainte ordinaires, dans lesquelles les contraintes correspondent à quelques mesures caractéristiques du fichier B (comme la moyenne ou la variance) auxquelles doivent satisfaire les variables du fichier qui a fait l'objet de l'appariement. De plus, les méthodes d'appariement sous contrainte ordinaires mettent l'accent sur la distribution marginale de Z , tandis que les méthodes étudiées ici mettent l'accent sur la distribution conditionnelle (bien qu'il s'agisse d'une distribution de variable nominale), qui est plus appropriée pour le fichier A; il existe donc une différence fondamentale entre les deux types de méthodes d'appariement sous contrainte. Conformément à Rubin (1986) et à Paass (1986), nous faisons une analyse empirique de l'efficacité des méthodes d'appariement. Nous avons donc effectué une étude de Monte Carlo pour analyser l'effet des modifications que l'on a proposé d'apporter aux méthodes existantes et ce, suivant deux scénarios: avec et sans information supplémentaire. Nous avons pu ainsi faire des analyses de sensibilité par rapport au non-respect de l'HIC et étudier les gains qui peuvent découler de l'utilisation d'informations supplémentaires. Les données fictives qui ont servi à l'étude de simulation ont été tirées de distributions normales multidimensionnelles accompagnées d'une contamination log-normale visant à induire l'asymétrie. L'utilisation de données fictives présente un avantage notable en ceci qu'on peut modifier les paramètres de contrainte pertinents pour obtenir différents scénarios de distribution pour le problème d'appariement. Nous avons comparé huit méthodes (quatre déjà en usage: REG, REG*, HOD, HOD*, et quatre à l'état de proposition: REG.LOGLIN, REG.LOGLIN*, HOD.LOGLIN, HOD.LOGLIN*) à l'aide de quatre mesures d'évaluation (deux au niveau individuel et deux au niveau global); se reporter à la section 6 pour plus de détails. Les principales conclusions de l'analyse empirique peuvent se résumer comme suit:

i) L'utilisation d'informations supplémentaires dans le but de contourner l'HIC peut améliorer considérablement la qualité du fichier enrichi. Toutefois, parmi les méthodes fondées sur l'HIC (c.-à-d. REG et HOD), la seconde est celle qui, dans l'ensemble, a la plus grande efficacité lorsqu'il n'existe pas d'information supplémentaire. En outre, on observe avec intérêt que

pratique, la méthode REG* ne sera pas applicable.

Par ailleurs, la méthode de Paass (dont une version est désignée dans cet article par le sigle HOD*) consiste essentiellement à déterminer tout d'abord une valeur intermédiaire, Z_{int} , à l'aide du fichier C en recourant à l'imputation hot-deck (avec une distance Y ou (X, Y) selon le cas), puis à déterminer une valeur Z authentique à partir du fichier B en utilisant de nouveau la méthode hot-deck avec une distance euclidienne (X, Z) . Il s'agit là d'une version simplifiée de la méthode originale de Paass, laquelle est itérative et consiste à mettre à jour successivement les valeurs de Z (pour le fichier A), de Y (pour le fichier B) et de X (pour le fichier C – en supposant que celui-ci ne contient que de l'information sur (Y, Z)) en utilisant les fichiers C, A et B respectivement jusqu'à ce qu'un critère de convergence soit satisfait; se reporter à la section 3 pour plus de détails. Pour démarrer l'itération, on impute convenablement les valeurs initiales de Z (pour A), de Y (pour B) et de X (pour C). Dans l'étude exposée ici, nous considérons uniquement la version simplifiée de la méthode de Paass en raison des calculs complexes qu'exige la méthode originale. Comme pour la méthode REG*, la méthode HOD* ne s'applique pas si le fichier C est sous forme de tableau de fréquences. En outre, même si le fichier C renferme des micro-données mais qu'il a une taille réduite (comme dans le cas d'une petite enquête menée à des fins particulières), ou lorsqu'il contient de l'information substantielle ou périmée, il peut être préférable d'extraire de l'information de macro-niveau, comme les distributions de variables nominales fondées sur une partition relativement grossière.

Il convient de souligner qu'en l'absence d'information supplémentaire, c.-à-d. d'un fichier C, les méthodes REG* et HOD* se ramènent simplement aux méthodes d'imputation ordinaires, notamment la méthode par régression (REG) et la méthode hot-deck (HOD). Ces méthodes font l'objet, elles aussi, de l'étude exposée ici.

Nous proposons des versions modifiées des méthodes de Rubin et de Paass désignées respectivement par REG.LOGLIN* et HOD.LOGLIN*; ces versions reposent sur la méthode d'imputation log-linéaire de Singh (1988). Selon ces versions modifiées, on se sert de l'information supplémentaire pour définir des contraintes nominales destinées aux fichiers appariés à l'aide des méthodes REG* et HOD*. De cette manière, les liens catégoriques (estimés par des modèles log-linéaires), qui induisent dans quelle mesure l'hypothèse d'indépendance conditionnelle n'est pas respectée (au sens qualitatif), se trouvent conservés dans le fichier qui a fait l'objet de l'appariement. Ces contraintes nominales sont censées rendre les distributions conjointes des données du fichier A robustes à

est une application importante de l'appariement statistique en analyse des lignes de conduite économiques (par ex. : calcul des impôts et des transferts pour les familles de la base de données). Dans la construction de la BDSPS, qui se fait en plusieurs étapes, on recourt à l'appariement statistique à certains moments pour enrichir le fichier receveur – l'enquête sur les finances des consommateurs (EFC) – de données d'autres sources. Plus précisément, on ajoute dans les enregistrements de l'EFC des données provenant des demandes de prestations d'assurance-chômage, des déclarations de revenus des particuliers et de l'enquête sur les dépenses des familles. Si le fichier A correspond à l'EFC et le fichier B, au fichier de données fiscales, alors X peut représenter des variables démographiques et économiques, Y peut désigner le revenu sous forme de transferts et Z peut représenter l'impôt à payer, le revenu de placements et les déductions.

Tel qu'il est décrit ci-dessus, l'appariement statistique présente une lacune grave en ceci qu'on ne connaît absolument rien de la variable Y. Cette lacune équivaut à l'hypothèse de l'indépendance conditionnelle de Y et Z étant donné $X (Y \perp Z | X)$, désignée par HIC (hypothèse d'indépendance conditionnelle). L'importance de cette hypothèse est indéniable puisque l'objectif de l'appariement est d'analyser les relations conjointes entre X, Y et Z. Si les relations réelles entre les variables sont telles que l'indépendance conditionnelle n'existe pas, l'HIC aura pour effet de dissimuler un aspect important de ces relations et de fausser des analyses qui font intervenir toutes les variables. Les risques de l'utilisation de l'HIC ont été soulignés par Sims (1978) et Rubin (1986), et bien que les appariements statistiques fondés sur l'HIC ne soient pas nécessairement entachés d'un biais important, Paass (1986) et Armstrong (1989) montrent, données empiriques à l'appui, que le problème est souvent réel. Si nous avons réalisé cette étude, c'est justement parce que nous cherchions à améliorer le contenu de la BDSPS, où l'appariement statistique repose sur l'HIC; voir aussi les commentaires de Schuren (1989) sur la méthodologie utilisée dans la BDSPS.

Les premiers ouvrages sur l'appariement statistique remontent à plus de vingt ans. Notons particulièrement ceux de Budd et Radner (1969), de Budd (1971) et d'Okner (1972). Dans ses observations sur l'article d'Okner, Sims (1972) est le premier qui souligne les risques potentiels de l'appariement statistique attribuables à l'existence de l'hypothèse d'indépendance conditionnelle. Fellegi (1977) s'interroge aussi sur la validité des distributions conjointes dans le fichier enrichi et il propose de soumettre les méthodes d'appariement à des tests empiriques complets. Le Département du commerce des États-Unis (U.S. Department of Commerce 1980) fait une bonne analyse des méthodes d'appariement statistique et des méthodes d'appariement exact; voir aussi Kadane (1978) et Rodgers (1984). Barr et Turner (1990) décrivent une étude empirique détaillée qui a pour objet la qualité des opérations dans la fusion de fichiers et présentent une bibliographie intéressante. Pour une analyse plus récente, le lecteur se référera à Cohen (1991).

Dans cet article, nous envisageons l'utilisation d'information supplémentaire en remplacement de l'HIC dans l'appariement statistique. Par conséquent, nous supposons l'existence d'un troisième fichier, C, qui contient de l'information supplémentaire sur l'ensemble complet de variables (X, Y, Z) ou l'ensemble réduit (Y, Z) . Il peut s'agir d'informations périmées ou substitutives (c.-à-d. variables différentes mais connexes) ou d'informations disposées en tableaux de fréquences, qui proviennent d'enquêtes spéciales ou de fichiers confidentiels. Nous voulons compléter les enregistrements du fichier A en allant chercher des valeurs de Z dans le fichier B sur la base d'informations contenues dans les fichiers A, B et C au sujet des relations conjointes entre X, Y et Z. On peut mesurer le succès de cette opération en évaluant la proportion des valeurs de Z contenues dans le fichier A qui proviennent vraisemblablement de la distribution réelle subordonnée à X et à Y. Dans une étude de simulation, nous pouvons comparer les valeurs appariées de Z aux valeurs vraies de Z qui ont été supprimées en nous servant de mesures d'évaluation au niveau individuel ou au niveau global. L'écart absolu moyen par rapport à la valeur vraie de Z et l'écart entre la covariance conditionnelle, $Cov(Y, Z | X)$, et la valeur vraie correspondante sont des exemples de mesures d'évaluation au niveau individuel. En ce qui concerne les mesures d'évaluation au niveau global, mentionnons la distance du chi carré et les valeurs P fondées sur les tests du rapport de vraisemblance pour distributions de variables nominales. Il arrive souvent dans la pratique que le fichier A, une fois complet, serve à construire des tableaux de recoupement; dans ce cas, l'attention se portera surtout sur les mesures de niveau global fondées sur des distributions de variables nominales. En outre, pour n'importe quelle distribution arbitraire de (X, Y, Z) , qui peut être fort complexe dans la pratique, la transformation en variable nominale est une méthode simple et uniforme pour résumer la distribution conjointe.

Comme nous l'avons mentionné plus haut, le problème de l'appariement statistique à une imputation indéfinie au point de vue pratique. En effet, pour un problème donné, on devrait choisir la méthode d'appariement en fonction du genre d'information supplémentaire dont on dispose. Le gros des méthodes proposées dans les ouvrages antérieurs est attribuable à Rubin (1986) et à Paass (1986). Le premier a proposé des méthodes de régression paramétrique tandis que le second a proposé des méthodes de régression non paramétrique. Les deux types de méthodes ont traité respectivement aux méthodes d'imputation par régression (REG) et aux méthodes d'imputation hot-deck (HOD), toutes bien connues.

La méthode de Rubin (dont une version est désignée dans cet article par le sigle REG*) consiste essentiellement à déduire tout d'abord une valeur intermédiaire, Z_{int} , du prédicteur par régression de Z par rapport à X et à Y (ce prédicteur étant lui-même obtenu à l'aide d'information sur la corrélation inconditionnelle $p_{Y,Z}$ ou la corrélation conditionnelle $p_{Y,Z|X}$ dans le fichier C), puis à déterminer une valeur Z authentique à partir du fichier B en utilisant la méthode hot-deck avec une distance euclidienne (X, Z) :

Appariement statistique: l'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle

A.C. SINGH, H.J. MANTTEL, M.D. KINACK et G. ROWE¹

RÉSUMÉ

Lorsqu'on crée des bases de données de microsimulation, souvent utilisées dans la planification et l'analyse des politiques, on combine plusieurs fichiers de données par des techniques d'appariement statistique afin d'enrichir le fichier receveur. Or, pour effectuer cette opération, il faut poser l'hypothèse de l'indépendance conditionnelle (HIC), ce qui peut fausser sérieusement les relations conjointes entre les variables. Dans cet article, nous examinons des méthodes hypothèse en utilisant des informations supplémentaires appropriées. On peut éviter de poser cette hypothèse en utilisant des méthodes d'imputation – par régression, hot-deck et log-linéaire – appliquées suivant deux scénarios: avec et sans information supplémentaire. La méthode d'imputation log-linéaire consiste essentiellement à introduire des contraintes nominales dans la méthode hot-deck. À partir d'une vaste étude de simulation faite avec des données fictives, nous exécutons des analyses de sensibilité lorsque l'on s'éloigne de l'HIC et nous étudions les gains qui peuvent découler de l'utilisation d'informations supplémentaires. À l'aide de données fictives, nous créons différents scénarios relatifs à la distribution et aux relations des variables pertinentes, par exemple distribution symétrique vs. distribution asymétrique et données supplémentaires substitutives vs. données supplémentaires non substitutives. Nous faisons aussi quelques recommandations sur l'utilisation des méthodes d'appariement statistique. Notre étude confirme particulièrement que l'HIC peut représenter une contrainte sérieuse, que l'on peut éliminer en utilisant des informations supplémentaires appropriées. L'étude montre aussi que les méthodes hot-deck sont généralement préférables aux méthodes de régression. De plus, lorsqu'on dispose d'informations supplémentaires, les contraintes nominales log-linéaires peuvent accroître l'efficacité des méthodes hot-deck. L'idée de cette étude est née des préoccupations que l'on avait sur l'utilisation de l'HIC dans la construction de la Base de données de simulation des politiques sociales à Statistique Canada.

MOTS CLÉS: Contraintes nominales; corrélation conditionnelle; contaminations log-normales; réduction à la moyenne.

1. INTRODUCTION

On peut envisager l'appariement statistique comme un cas particulier de l'imputation; celle-ci implique deux sources de micro-données qui contiennent des renseignements différents sur des unités différentes. L'une de ces sources sert de fichier cible, ou fichier receveur, dans lequel des données sont imputées pour chaque enregistréement à partir des données de l'autre source, qui est définie comme le fichier donneur. Cependant, ce rapport entre l'appariement statistique et l'imputation disparaît dès lors que le fichier receveur contient des renseignements sur des variables que l'on ne retrouve pas dans le fichier donneur. Le fichier apparié (enrichi) sert le plus souvent de source de paramètres pour les modèles de micro-simulation, qui exigent normalement un fichier complet, comprenant toutes les variables. Les fichiers de micro-données connus peuvent équivaloir à des échantillons tirés de fichiers administratifs ou d'enquêtes. Comme les enregistréements tirés de divers fichiers sources se rapportent à des unités différentes, le processus de fusion des données des divers fichiers diffère de l'appariement exact, où l'on recherche plutôt des unités précises

dans les diverses sources. En fait, même s'il était possible de faire un appariement exact des fichiers, des règles de protection du secret statistique viendraient imposer certaines contraintes à ce processus. De façon générale, le problème se formule comme suit. Un fichier receveur A contient de l'information sur les variables (X, Y) et un fichier donneur B, de l'information sur les variables (X, Z) . La variable commune aux deux fichiers, X , peut servir à reconnaître les unités identiques dans les deux fichiers. Le problème consiste à compléter les enregistréements du fichier A en imputant des valeurs authentiques (observées) pour Z à l'aide de l'information du fichier B sur la relation (X, Z) . Dans la pratique, les variables X, Y et Z sont généralement multidimensionnelles. Un précieux avantage de l'imputation de valeurs authentiques de Z est de préserver les relations entre les éléments de la variable multidimensionnelle Z . Pour des raisons de commodité, nous supposons dans cet article que X, Y et Z sont des variables unidimensionnelles. La Base de données de simulation des politiques sociales (BDSPS; voir Wolfson *et al.*, 1987), qui est une base de données de microsimulation créée à Statistique Canada,

¹ A.C. Singh, H.J. Mantel et M.D. Kinack, Division des méthodes d'enquêtes sociales; G. Rowe, Division des études sociales et économiques, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

BIBLIOGRAPHIE

- BEEBE, G. W. (1985). Why are epidemiologists interested in matching algorithms? Dans *Record Linkage Techniques* - 1985. U.S. Internal Revenue Service.
- BELIN, T. (1991). Using Mixture Models to Calibrate Error Rates in Record Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation. Harvard Ph.D. Thesis.
- CARPENTER, M., et FAIR, M.E. (Éditeurs) (1989). *Proceedings of the Record Linkage Sessions and Workshops*, Canadian Epidemiological Research Conference, Statistique Canada.
- COOMBS, J.W., et SINGH, M.P. (Éditeurs) (1987). *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada.
- COPAS, J.B., et HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society A*, 153, 287-320.
- CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, D.B. (1992). Evaluation of a new procedure for estimating income and tax aggregates from advance data. *Journal of Business and Economic Statistics*, 10, 117-131.
- DRAPER, N.R., et SMITH, H. (1981). *Applied Regression Analysis*, 2ième édition. New York: J. Wiley.
- FELTEGI, I., et SUNTER, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.
- HOWE, G., et SPASOFF, R.A. (Éditeurs) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto, Ontario, Canada: University of Toronto Press.
- JABINE, T.B., et SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.
- JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- JOHNSTON, J. (1972). *Econometric Methods*, 2ième Edition. New York: McGraw-Hill.
- KILSS, B., et ALVEY, W. (Éditeurs) (1985). *Record Linkage Techniques* - 1985. U.S. Internal Revenue Service, Publication 1299, 2-86.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies*, Administration and Business. Oxford: Oxford University Press.
- NEWCOMBE, H.B., FAIR, M.E., et LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1208.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NETER, J., MAYNES, E.S., et RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- ROSENBAUM, P., et RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- ROSENBAUM, P., et RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.
- RUBIN, D.B. (1990). Discussion (of Imputation Session). *Proceedings of the Bureau of the 1990 Annual Research Conference*, U.S. Bureau of the Census, 676-678.
- RUBIN, D., et BELIN, T. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.
- SCHEUREN, F. (1985). Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques* - 1985. U.S. Internal Revenue Service.
- SCHEUREN, F., et OH, H.L. (1975). Fiddling Around with Nonmatches and Mismatches. *Proceedings of the Social Statistics Section, American Statistical Association*, 627-633.
- SCHEUREN, F., OH, H.L., VOGEL, L., et YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages, U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.
- TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- WINKLER, W.E. (1985). Exact matching list of businesses: blocking, subfile identification, and information theory. Dans *Record Linkage Techniques* - 1985. (Eds. B. Kilss et W. Alvey). U.S. Internal Revenue Service, Publication 1299, 2-86.
- WINKLER, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.
- WINKLER, W.E., et SCHEUREN, F. (1991). How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis. U.S. Bureau of the Census, Statistical Research Division Technical Report.
- WINKLER, W.E., et THIBAUDEAU, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical report.

A.4. Modèle de Rubin-Belin

Pour estimer la probabilité d'un lien vrai dans n importe quel intervalle de poids, Rubin et Belin (1991) considèrent l'ensemble de paires produites par le programme de couplage par ordinateur, rangées par ordre décroissant de poids. Ils supposent que la probabilité d'un lien vrai est une fonction monotone du poids; autrement dit, plus le poids est élevé, plus grande est la probabilité d'un lien vrai. Ils supposent que la distribution des poids observés est une combinaison des distributions des liens vrais et des non-liens vrais.

Leur méthode d'estimation est la suivante:

1. Modéliser chacune des deux composantes de la combinaison comme une distribution normale avec moyenne et variance inconnues après des transformations puissance distinctes.

2. Estimer la puissance des deux transformations à l'aide d'un échantillon d'apprentissage.

3. À l'aide des deux transformations ainsi déterminées, ajuster un modèle de combinaison normale aux données, courantes sur les poids, pour obtenir des estimations du maximum de vraisemblance (et des erreurs-types).

4. Utiliser les paramètres du modèle ajusté pour obtenir des estimations ponctuelles du taux de liens faux en fonction du niveau seuil et obtenir les erreurs-types visant le taux de liens faux au moyen d'une approximation par la méthode delta.

Bien que la méthode de Rubin-Belin exige un échantillon d'apprentissage, celui-ci sert surtout à obtenir la forme des courbes. Autrement dit, si la transformation puissance est donnée par

$$\psi(w_i; \delta, \omega) = \begin{cases} \omega \log(w_i) & \text{si } \delta = 0, \\ (w_i^\delta - 1) / (\delta \omega^{\delta-1}) & \text{si } \delta \neq 0 \end{cases}$$

où ω est la moyenne géométrique des poids w_i , $i = 1, \dots, n$, alors ω et δ peuvent être estimés pour les deux courbes. Dans le cas des exemples du présent article et d'une vaste gamme d'autres situations de couplage (Winkler et Thibaudau 1991), la méthode d'estimation de Rubin-Belin donne de bons résultats. Dans certaines autres situations, une méthode différente (Winkler 1992) utilisant plus d'information que la méthode de Rubin-Belin et n'exigeant pas d'échantillon d'apprentissage donne des estimations exactes, tandis que le logiciel (voir p. ex. Belin 1991) basé sur la méthode de Rubin-Belin ne donne pas d'estimation convergente, même si de nouvelles données d'étalement sont obtenues. Puisque les données d'étalement visant les scénarios efficace et médiocre du présent article sont appropriées, la méthode de Rubin-Belin donne de meilleures estimations que la méthode de Winkler.

$$(n - p) \hat{\sigma}^2 = (Y - X\beta)^T (Y - X\beta)$$

$$= Y^T Y - \hat{\beta} X^T Y, \quad (\text{A.2.4})$$

$$\text{ou } \hat{\beta} = (X^T X)^{-1} X^T Y.$$

En utilisant (A.2.1), on peut représenter $\hat{\beta} X^T Y$ en termes des variables observables Z et X , de la même façon qu'en (A.1.2) et en (A.1.3). Puisque

$$Y^T Y = Z^T Z - B^T Z + Z^T B + B^T B, \quad (\text{A.2.5})$$

nous pouvons obtenir la portion restante du côté droit de (A.2.4) qui permet l'estimation de σ^2 . Selon la formule habituelle (p. ex. Graybill 1976, p. 276), la covariance de $\hat{\beta}$ est

$$\text{cov}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}, \quad (\text{A.2.6})$$

ce que nous pouvons estimer.

A.3. Régression multiple avec variables indépendantes venant des deux fichiers

Lorsque certaines des variables indépendantes viennent du même fichier que Y , nous devons les redresser d'une manière analogue à celle employée pour redresser X dans les équations (A.1.1) et (A.2.1). La matrice X peut alors s'écrire sous la forme

$$X_d = X + D, \quad (\text{A.3.1})$$

où D est la matrice des redressements tenant compte du biais qui redonne aux termes de X venant du même fichier que Y leurs valeurs vraies qui sont représentées dans X_d . En utilisant (A.2.1) et (A.2.2), nous obtenons

$$Y + B = (X_d - D)C. \quad (\text{A.3.2})$$

Après quelques opérations algébriques, (A.3.2) devient

$$(X_d^T X_d)^{-1} X_d^T Y = (X_d^T X_d)^{-1} X_d^T (-B)$$

$$+ (X_d^T X_d)^{-1} X_d^T (X_d + D)C$$

$$= (X_d^T X_d)^{-1} X_d^T (-B)$$

$$+ (X_d^T X_d)^{-1} X_d^T DC + C. \quad (\text{A.3.3})$$

Si D est zéro (c.-à-d. que toutes les valeurs x indépendantes proviennent d'un même fichier), (A.3.3) correspond à (A.2.3). Le premier terme du côté gauche de (A.2.3) est l'estimation de $\hat{\beta}$. L'estimation $\hat{\sigma}^2$ est obtenue d'une manière analogue à la façon dont (A.2.3), (A.2.4) et (A.2.5) ont été utilisés. La covariance de $\hat{\beta}$ découle de (A.2.6).

$$\sigma_{zx} \equiv E[(Z - EZ)(X - EX)]$$

$$= (1/n) \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) p_i$$

$$+ (1/n) \sum_{j \neq i} (Y_j - \bar{Y})(X_i - \bar{X}) q_{ij}$$

$$= (1/n) S_{yx} + B_{yx} = \sigma_{yx} + B_{yx}, \quad (A.1.3)$$

où $B_{yx} = (1/n) \sum_i [(Y_i - \bar{Y})(X_i - \bar{X}) (-h_i) + (Y_{\phi(i)} - \bar{Y})(X_i - \bar{X}) h_i]$, $S_{yx} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$ et $\sigma_{yx} = (1/n) S_{yx}$. Le terme B_{yx} est le biais pour les deuxièmes moments et le terme B_{yx} est le biais pour le produit croisé de Y et de X . Les formules (A.1.1), (A.1.2) et (A.1.3) correspondent respectivement aux formules (A.1), (A.2) et (A.3) de Neter et coll. Les formules diffèrent nécessairement dans leurs détails, parce que nous utilisons un modèle plus général du processus de couplage.

Pour les coefficients de régression, on a

$$\beta_{zx} \equiv \sigma_{zx}/\sigma_x^2 = \sigma_{yx}/\sigma_x^2 + B_{yx}/\sigma_x^2 = \beta_{yx} + B_{yx}/\sigma_x^2. \quad (A.1.4)$$

Pour obtenir une estimation de la variance de β_{yx} , nous estimons d'abord s^2 pour obtenir la variance σ^2 de la façon habituelle.

$$(n - 2) s^2 = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} + \beta_{yx} S_{xy}$$

$$= n \sigma_y^2 - n \beta_{yx} \sigma_x^2. \quad (A.1.5)$$

L'utilisation de (A.1.2) et de (A.1.3) nous permet d'exprimer s^2 en termes des quantités observables σ_y^2 et σ_{zx} et des termes de biais B_{yy} , B_{yx} et B qui peuvent être calculés en vertu de nos hypothèses. La variance estimée de β_{yx} est alors calculée au moyen de la formule habituelle (p. ex. Draper et Smith 1981, p. 18-20)

$$\text{Var}(\beta_{yx}) = s^2 / (n \sigma_x^2).$$

Nous observons que la première égalité dans (A.1.5) fait intervenir l'hypothèse habituelle des régressions selon laquelle les termes d'erreur sont indépendants et ont une variance identique.

Dans les exemples numériques du présent article, nous

avons supposé que la valeur indépendante vraie X_i associée à chaque X_i provenait de l'enregistrement ayant le poids d'appariement le plus élevé et que la valeur indépendante fautive provenait de l'enregistrement ayant le poids d'appariement venant au deuxième rang. Cette hypothèse est plausible, car nous n'avons examiné que la régression simple dans le présent article, et parce que le poids d'appariement venant au deuxième rang était généralement beaucoup plus faible que le poids le plus élevé. Par conséquent, il est beaucoup plus naturel de supposer que l'enregistrement dont le poids vient au deuxième rang donne la valeur fautive. Dans nos exemples empiriques, nous effectuons des redressements directs et posons des hypothèses simples

qui donnent de bons résultats, car ils sont conformes à la nature des données et au processus de couplage. Dans des situations de régression plus complexes, ou avec d'autres modèles comme les modèles log-linéaires, il nous faudra vraisemblablement poser d'autres hypothèses. Ce besoin d'hypothèses additionnelles s'apparente au cas des modèles simples de non-réponse, qui exigent des hypothèses additionnelles lorsqu'on passe de la non-réponse dont on n'a pas à tenir compte à la non-réponse dont il faut tenir compte (voir Rubin 1987).

A.2. Régression multiple avec variables dépendantes venant d'un fichier et variables dépendantes venant de l'autre fichier

À ce stade, nous passons à la notation matricielle courante (p. ex. Graybill 1976). Notre modèle de base est

$$Y = XB + \epsilon,$$

où Y est une matrice $n \times 1$, X est une matrice $n \times p$, B est une matrice $p \times 1$ et ϵ est une matrice $n \times 1$. En suivant le même raisonnement qu'en (A.1.1), nous pouvons écrire

$$Z = Y + B, \quad (A.2.1)$$

où Z , Y et B sont des matrices $n \times 1$ ayant des termes liés entre eux, pour $i = 1, \dots, n$, par

$$z_i = y_i + p_i y_i + h_i y_{\phi(i)}.$$

Parce que nous observons Z et X seulement, nous considérons l'équation

$$Z = XC + \epsilon. \quad (A.2.2)$$

Nous obtenons une estimation \hat{C} en effectuant la régression par rapport aux données observées de la façon habituelle. Nous voulons redresser l'estimation \hat{C} pour obtenir une estimation $\hat{\beta}$ de β d'une manière analogue à (A.1.1). En utilisant (A.2.1) et (A.2.2), nous obtenons

$$(X^T X)^{-1} X^T Y + (X^T X)^{-1} X B = \hat{C}. \quad (A.2.3)$$

Le premier terme du côté gauche de (A.2.3) est l'estimation habituelle $\hat{\beta}$. Le deuxième terme du côté gauche de (A.2.3) est la matrice transposée de X . X^T est la matrice transposée de X . La formule habituelle (Graybill 1976, p. 176) permet l'estimation de la variance σ^2 associée aux composantes d'erreur indépendantes et identiquement distribuées de ϵ

REMERCIEMENTS ET RESPONSABILITES

Les auteurs voudraient remercier Yahia Ahmed et Mary Batcher pour leur contribution à la préparation de cet article, ainsi que deux arbitres pour leurs commentaires détaillés et pertinents. Des échanges fructueux ont eu lieu avec Tom Belin. Wendy Alvey nous a également fourni une aide importante sur le plan de la rédaction.

Selon la coutume, il importe de préciser que cet article reflète les opinions des auteurs et pas nécessairement celles des organismes dont ils font partie. Tout problème, par exemple un manque de clarté dans les idées ou les sujets présentés, est entièrement attribuable aux auteurs.

ANNEXE

L'annexe est divisée en quatre sections. La première explique de façon détaillée comment l'erreur de couplage se répercute sur les modèles de régression dans le cas unidimensionnel simple. L'approche est très voisine de celle présentée par Neter et coll. (1965) et sert d'inspiration aux généralisations présentées à la deuxième et à la troisième section de l'annexe. Les formules de calcul sont beaucoup plus complexes que celles présentées par Neter et coll. Puisque nous utilisons un modèle plus réaliste du processus de couplage. Dans la deuxième section, nous étendons le modèle unidimensionnel au cas où toutes les variables indépendantes proviennent d'un fichier, tandis que la variable dépendante vient de l'autre, et dans la troisième section, nous étendons le deuxième cas à celui où certaines variables indépendantes viennent d'un fichier et certaines viennent d'un autre. La quatrième section résume les méthodes de Rubin et Belin (1991) (voir aussi Belin 1991) permettant l'estimation de la probabilité d'un lien.

A.1. Modèle de régression unidimensionnel

Dans la présente section, nous examinons la situation de régression la plus simple, dans laquelle nous coupons deux fichiers et examinons un ensemble de paires numériques dans lequel la variable indépendante est tirée d'un enregistrement d'un fichier et la variable dépendante, de l'enregistrement correspondant de l'autre fichier.

Soit $Y = X\beta + \epsilon$ le modèle de régression unidimensionnel ordinaire pour lequel les termes d'erreur sont indépendants avec espérance zéro et variance constante σ^2 . Si nous examinons une seule base de données, la régression serait faite de Y en X de la façon habituelle. Pour $i = 1, \dots, n$, nous souhaitons utiliser (X_i, Y_i) , mais nous utilisons plutôt (X_i, Z_i) , où Z_i est habituellement Y_i mais pourrait prendre une autre valeur, Y_j , en raison de l'erreur de couplage.

Autrement dit, pour $i = 1, \dots, n$,

$$Z_i = \begin{cases} Y_i & \text{avec probabilité } p_i \\ Y_j & \text{avec probabilité } q_{ij} \text{ pour } j \neq i, \end{cases}$$

$$\text{où } p_i + \sum_{j \neq i} q_{ij} = 1.$$

La première et la deuxième égalité découlent des définitions et la troisième s'obtient par addition et soustraction. C'est à la troisième inégalité que nous appliquons pour la première fois l'hypothèse d'un couplage biunivoque. Le dernier terme du côté droit de l'égalité est le biais, que nous désignons par B . Notons que le biais global B est la moyenne statistique (espérance) des biais individuels $[Y_i (-h_i) + X_{\phi(i)} h_i]$ pour $i = 1, \dots, n$. De même, nous avons

$$\begin{aligned} E(Z) &= (1/n) \sum_i E(Z|i) = (1/n) \sum_i (Y_i p_i + \sum_{j \neq i} Y_j q_{ij}) \\ &= (1/n) \sum_i [Y_i (-h_i) + X_{\phi(i)} h_i] + Y + B. \end{aligned} \quad (\text{A.1.1})$$

La probabilité p_i peut être zéro ou un. Nous définissons $h_i = 1 - p_i$. Comme dans Neter et coll. (1965), nous divisons l'ensemble de paires en n classes mutuellement exclusives. Chaque classe est formée d'exactement un (X_i, Z_i) , de sorte qu'il y a n classes. L'idée intuitive de notre méthode est que nous redressons Z_i dans chaque (X_i, Z_i) pour tenir compte du biais introduit par le processus de couplage. L'exactitude du redressement est fortement liée à l'exactitude des estimations des probabilités de concordance dans notre modèle.

Pour simplifier les formules de calcul, nous supposons un couplage biunivoque, c.-à-d. que pour chaque $i = 1, \dots, n$, il existe au plus un j tel que $q_{ij} > 0$. Nous définissons ϕ par $\phi(i) = j$. Notre modèle demeure valable même si nous ne faisons pas l'hypothèse d'un couplage biunivoque.

À titre d'étapes intermédiaires dans l'estimation des coefficients de régression et de leurs erreurs-types, nous devons trouver $\mu_z = E(Z)$, σ_z^2 et σ_{xz} . Comme dans Neter et coll. (1965),

$$\begin{aligned} \sigma_z^2 &= E(Z - EZ)^2 = E(Z - Y + B)^2 \\ &= (1/n) \sum_i (Y_i - Y)^2 p_i + (1/n) \sum_{j \neq i} (Y_j - Y)^2 q_{ij} - 2BE(Z - Y) + B^2 \\ &= (1/n) S_{yy} + B_{yy} - B^2 = \sigma_y^2 + B_{yy} - B^2, \end{aligned}$$

$$\text{où } B_{yy} = (1/n) \sum_i [(X_i - \bar{X})^2 (-h_i) + (X_{\phi(i)})^2 h_i], \quad S_{yy} = \sum_i (Y_i - \bar{Y})^2 \text{ et } \sigma_y^2 = (1/n) S_{yy}. \quad (\text{A.1.2})$$

le fait que les classes de poids étaient de longueur égale et nous n'avons pas non plus étudié ce qui serait survenu si elles avaient été de longueurs différentes.

Dernière observation: comme nous l'avons déjà indiqué, nous croyons que notre approche a beaucoup en commun avec la méthode des scores de propension; toutefois, nous n'avons pas fait appel explicitement à cette théorie plus générale, mais nous croyons qu'il vaudrait la peine de le faire. Par exemple, les notions propres à cette théorie pourraient être particulièrement utiles dans le cas où les variables de régression et les variables de couplage sont dépendantes. (Voir Winkler et Schuren (1991) pour un rapport sur les simulations limitées qui ont été effectuées et les difficultés additionnelles rencontrées.)

5.2 Traitement des non-liens erronés

Dans l'application des méthodes de couplage d'entregistremments, le problème général du biais de sélection survient en raison de la présence de non-liens erronés. Il existe plusieurs façons de traiter ce problème. Par exemple, les liens pourraient être redressés par l'analyste pour tenir compte du manque de représentativité, au moyen de méthodes familières aux responsables des redressements effectués au titre de la non-réponse d'unités ou, peut-être même, de la non-réponse à des questions particulières (p. ex. Schuren et coll. 1981).

La présente méthode de traitement des liens possibles pourrait aider à réduire l'ampleur du problème des non-liens erronés, mais, en général, ne l'éliminerait pas. Plus précisément, supposons que nous ayons un cadre de couplage dans lequel, en raison d'un manque de ressources, il est impossible de faire le suivi des liens possibles. De nombreux praticiens pourraient simplement laisser de côté les liens possibles, accroissant du même coup le nombre de non-liens erronés. (Par exemple, dans la détermination des membres d'une cohorte qui sont vivants ou décédés, une troisième possibilité – indéterminé – est souvent utilisée.)

Notre approche à l'égard des liens possibles aurait *implicitement* produit un redressement pour la partie des non-liens erronés qui pouvaient éventuellement donner un lien (avec, disons, une étape de suivi). Les autres non-liens erronés demeureraient généralement irrésolus et la possibilité d'effectuer un autre redressement à leur sujet pourrait être une question à envisager.

Il se produit fréquemment des situations où les fichiers couplés comportent des sous-groupes qui recèlent une information de couplage de qualité variable, et qui présentent par conséquent des taux de liens et de non-liens erronés qui diffèrent. En principe, nous pourrions appliquer les techniques décrites ici à chacun de ces sous-groupes séparément. La façon de procéder face à des sous-groupes de très petite taille est un problème toujours sans réponse, et l'effet sur les différences estimées d'un sous-groupe à l'autre, même quand les deux sont de taille modeste, bien qu'il puisse sembler évident, mérite d'être étudié.

5.3 Derniers commentaires

Au début du présent article, nous avons posé deux questions "clés". Maintenant que nous arrivons à la fin, il est utile de revoir ces questions et de tenter, de façon sommaire, d'y répondre.

- *"Que peut faire le responsable du couplage pour aider l'analyste?"* Si possible, le responsable du couplage devrait jouer un rôle dans la conception des ensembles de données à coupler, de façon que l'information d'identification des deux ensembles soit de haute qualité. Plusieurs organismes disposent déjà d'algorismes puissants, capables d'effectuer un excellent couplage (c'est le cas à Statistique Canada et au U.S. Bureau of the Census, pour n'en nommer que deux). Les responsables du couplage devraient résister à la tentation de concevoir et d'élaborer leurs propres logiciels. Il est hautement recommandé, dans la plupart des cas, de modifier ou simplement d'utiliser les logiciels existants (Schuren 1985). De toute évidence, dans l'intérêt de l'analyste, le responsable du couplage doit fournir autant d'information que possible au sujet des fichiers couplés, afin que l'analyste puisse faire des choix éclairés dans son travail. Dans le présent article, nous avons proposé que les liens, les non-liens et les liens possibles soient fournis à l'analyste – et non pas uniquement les liens. Nous recommandons fortement qu'il en soit ainsi, même si une étape de révision manuelle a été réalisée. Nous ne recommandons *pas* nécessairement les choix particuliers que nous avons faits au sujet de la structure des fichiers, du moins tant qu'ils n'auront pas été étudiés davantage. Nous faisons valoir, toutefois, que nos choix présentent une certaine utilité.

- *"Que devrait savoir l'analyste au sujet du couplage et comment cette information devrait-elle être utilisée?"* L'analyste doit disposer de l'information sur les liens, les non-liens et les liens possibles, ainsi que les probabilités de concordance, si ces données sont disponibles. Il se peut que dans de nombreuses situations, la simple exécution des étapes d'analyse des données séparément pour chaque état de concordance permette d'en apprendre beaucoup sur la sensibilité des résultats. Le présent article énonce des idées initiales sur la façon dont cela peut se faire dans un contexte de régression. Il semble également que des améliorations soient possibles grâce aux redressements effectués ici, notamment dans le scénario de couplage médiocre. La mesure dans laquelle ces améliorations sont générales reste à vérifier. Malgré tout, nous sommes assez satisfaits de nos résultats et nous sommes impatients de pousser le sujet plus à fond. En effet, il y a des liens directs à établir entre notre approche du problème de la régression et d'autres techniques courantes, comme les modèles log-linéaires à tableaux de contingence.

De toute évidence, nous n'avons pas proposé de réponses complètes et générales aux questions soulevées. Nous espérons, toutefois, que cet article pourra au moins stimuler l'intérêt et susciter d'autres contributions qui nous aideront à améliorer nos méthodes.

Sommaire des résultats des redressements pour des simulations représentatives

Base des redressements	Scénario de couplage		
	Efficace	Médiocre	Pauvre
Probabilités vraies	Le redressement n'a pas été utile parce que non nécessaire	Bons résultats comme ceux de la section 4.1	Bons résultats comme ceux de la section 4.1
Probabilités estimées	Même commentaire que ci-dessus	Même commentaire que ci-dessus	Piètres résultats, car la méthode de Rubin-Belin ne permettait pas l'estimation des probabilités

Il sera difficile, quelle que soit la méthode d'estimation statistique, d'obtenir de bons résultats avec le scénario de couplage pauvre, en raison de l'énorme chevauchement des courbes. Voir la figure 4. Nous croyons que le scénario médiocre couvre une vaste gamme de cadres généralement rencontrés. Le scénario de couplage pauvre, toutefois, pourrait également survenir assez souvent, en particulier si les responsables du couplage sont moins expérimentés. Il faudra ou bien que de nouvelles méthodes d'estimation soient élaborées pour le scénario pauvre, ou bien que la méthode d'estimation des probabilités de Rubin-Belin – qui n'a pas été conçue pour cette situation – soit améliorée.

4.3 Limites des simulations

Les résultats des simulations souffrent d'un certain nombre de limites. Quelques-unes pourront avoir des répercussions pratiques importantes, d'autres moins. En voici une liste partielle:

- Dans l'exécution des simulations relatives au présent article, nous avons supposé que la paire ayant le poids le plus élevé était un lien vrai et que la paire dont le poids venait au deuxième rang était un non-lien vrai. Cette hypothèse n'est pas valable, car c'est parfois le poids venant au deuxième rang qui correspond à un lien vrai, et le poids le plus élevé qui correspond à un non-lien vrai. (Nous ne savons pas trop quelle pourrait être l'importance de ce fait en pratique. Il s'agirait sûrement d'un facteur important dans le cas des scénarios de couplage pauvres.)
- Une deuxième limite des ensembles de données ayant servi aux simulations vient de la possibilité que l'enregistrement réellement lié ne figure pas du tout dans le fichier auquel le premier fichier est couplé. (Ce pour-rat être un facteur important. Dans beaucoup de situations pratiques, il serait normal de s'attendre aussi à ce que les "critères de groupage logique" fussent en sorte

- Une troisième limite de notre approche est qu'aucune utilisation n'a été faite des outils diagnostiques classiques de la régression. (Selon l'environnement examiné, les valeurs aberrantes créées par la présence de non-liens pourraient bouleverser les relations sous-jacentes. Cela n'a pas représenté un gros problème dans nos simulations, en grande partie, peut-être, parce que les valeurs X et Y créées étaient confinées à un intervalle modérément étroit.)

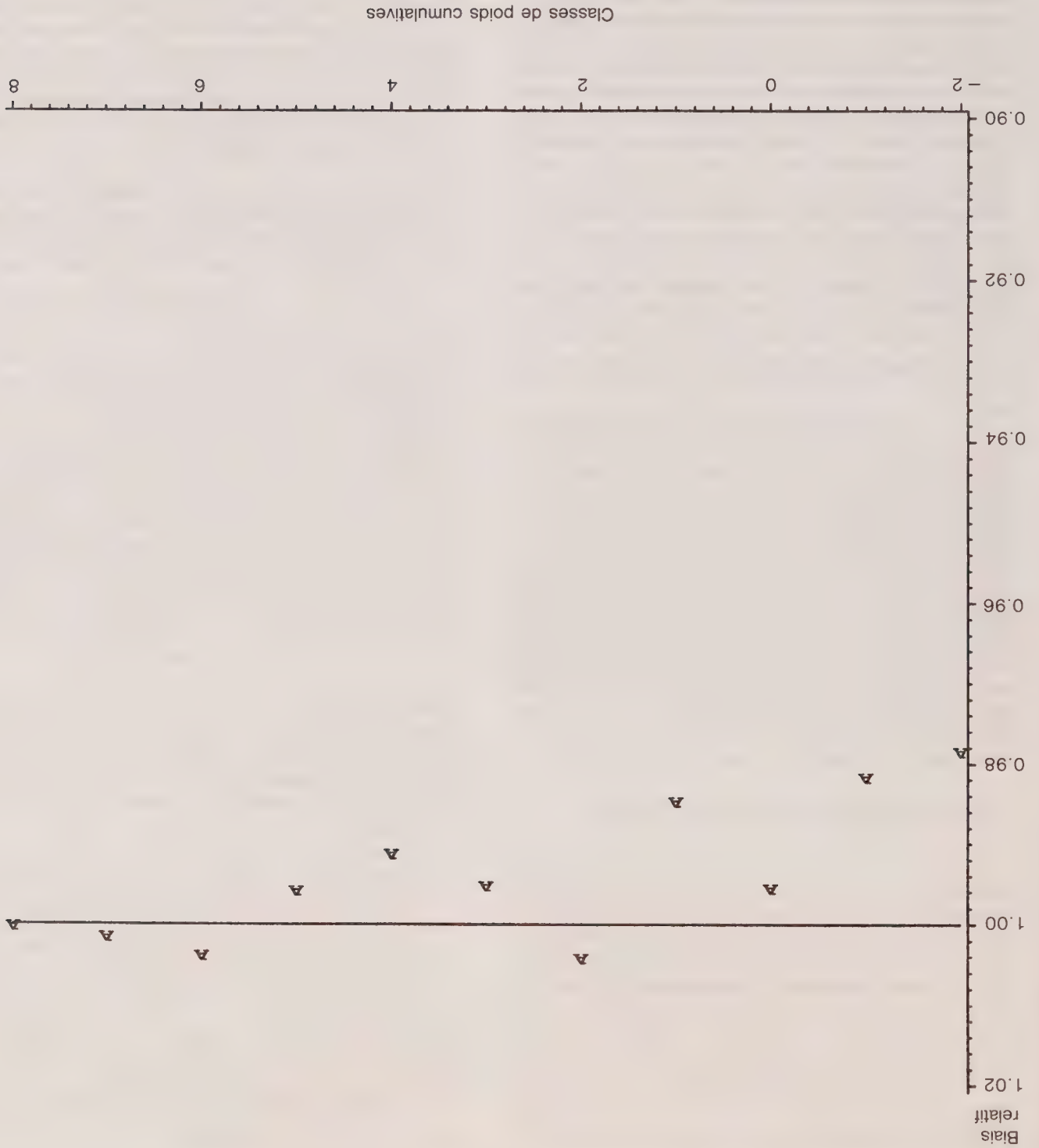
5. CONCLUSIONS ET TRAVAUX FUTURS

Les résultats théoriques et les données de simulation connexes présentés ici sont évidemment quelque peu fabriqués et artificiels. Il reste donc beaucoup à faire pour confirmer et généraliser nos efforts initiaux. On peut toutefois en tirer certaines recommandations valables pour les applications actuelles, ainsi que des suggestions de recherche future. Nous allons d'abord examiner quelques aspects qui ont attiré notre attention et qui mériteraient qu'on les étudie plus à fond afin d'améliorer les redressements touchant les liens possibles. Ensuite, nous formulerons certaines remarques sur le problème connexe du traitement à réserver aux non-liens (restants). Enfin, nous énoncerons certaines idées sommaires et nous rappellerons notre perspective quant à l'unité des tâches accomplies par les responsables du couplage et les analystes.

5.1 Améliorations de la méthode de redressement

Une question évidente consiste à déterminer si nos méthodes de redressement pourraient s'inspirer des méthodes générales concernant les erreurs sur les variables (p. ex. Johnson 1972). Nous n'avons pas exploré cet aspect, mais l'effort pourrait en valoir la peine. Les techniques qui émanent des outils diagnostiques classiques de la régression sont plus intéressantes pour nous. Une combinaison de ces dernières avec notre approche pourrait se révéler très intéressante. Rappelons que nous effectuons des redressements s'appliquant tout à tour à chacune des classes de poids. Supposons que nous devions l'examen du diagramme de dispersion résiduel d'une classe de poids particulièrement, dans laquelle les résidus ont été calculés autour de la régression obtenue d'après les classes de poids cumulatives situées au-dessus de la classe en question. Des valeurs aberrantes pourraient alors être identifiées et traitées comme des non-liens plutôt que comme des liens possibles.

Nous nous proposons d'explorer cette possibilité à l'aide de données plus concentrées aux extrêmes que celles que nous avons utilisées ici. En outre, nous examinerons les résultats obtenus en faisant varier la longueur des classes de poids et le nombre minimum de cas dans chaque classe. Nous craignons en effet que le nombre de cas de chaque classe ait été trop faible dans certains cas. (Voir le tableau 1.) Par contre, nous n'avons pas exploité



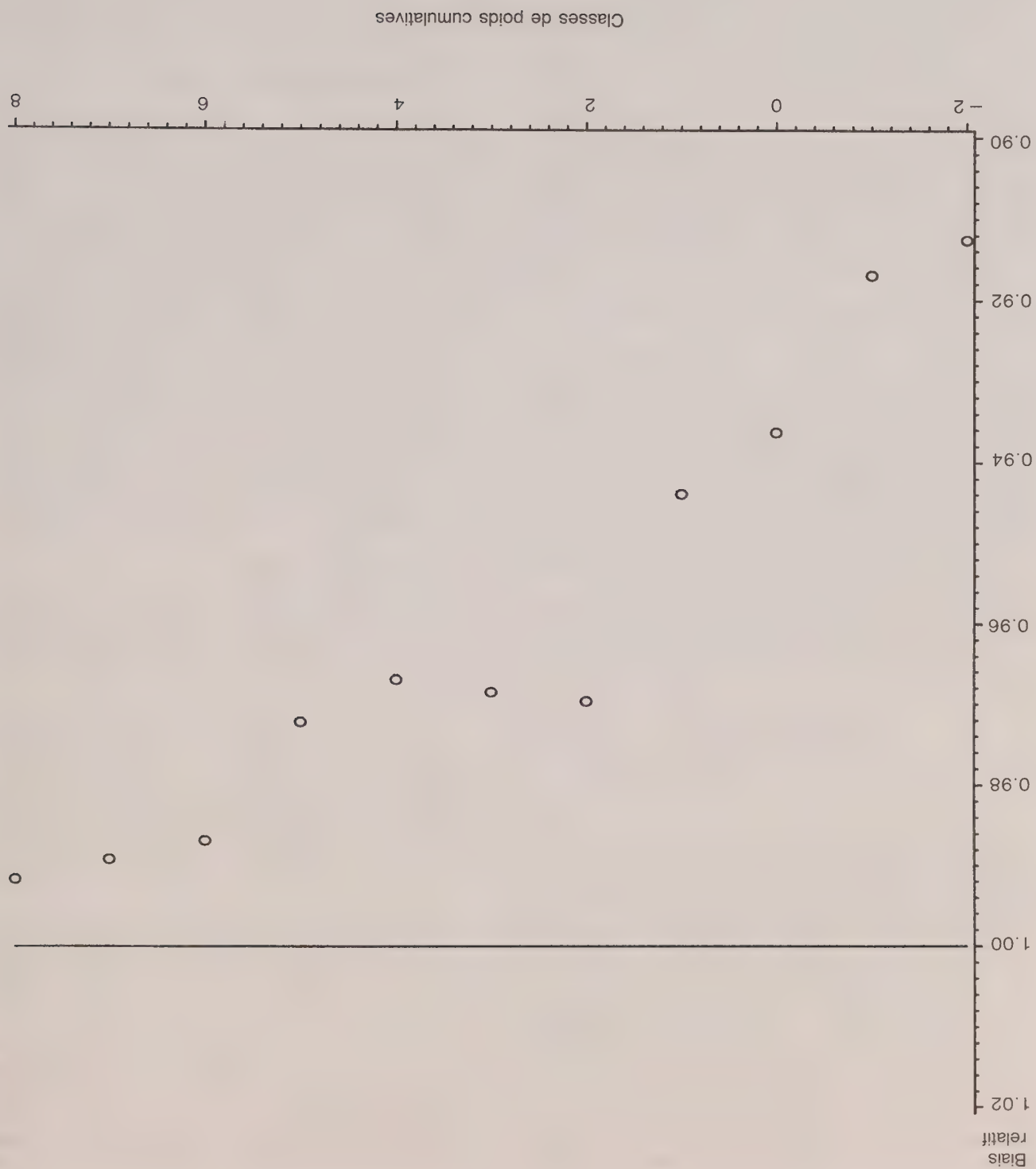
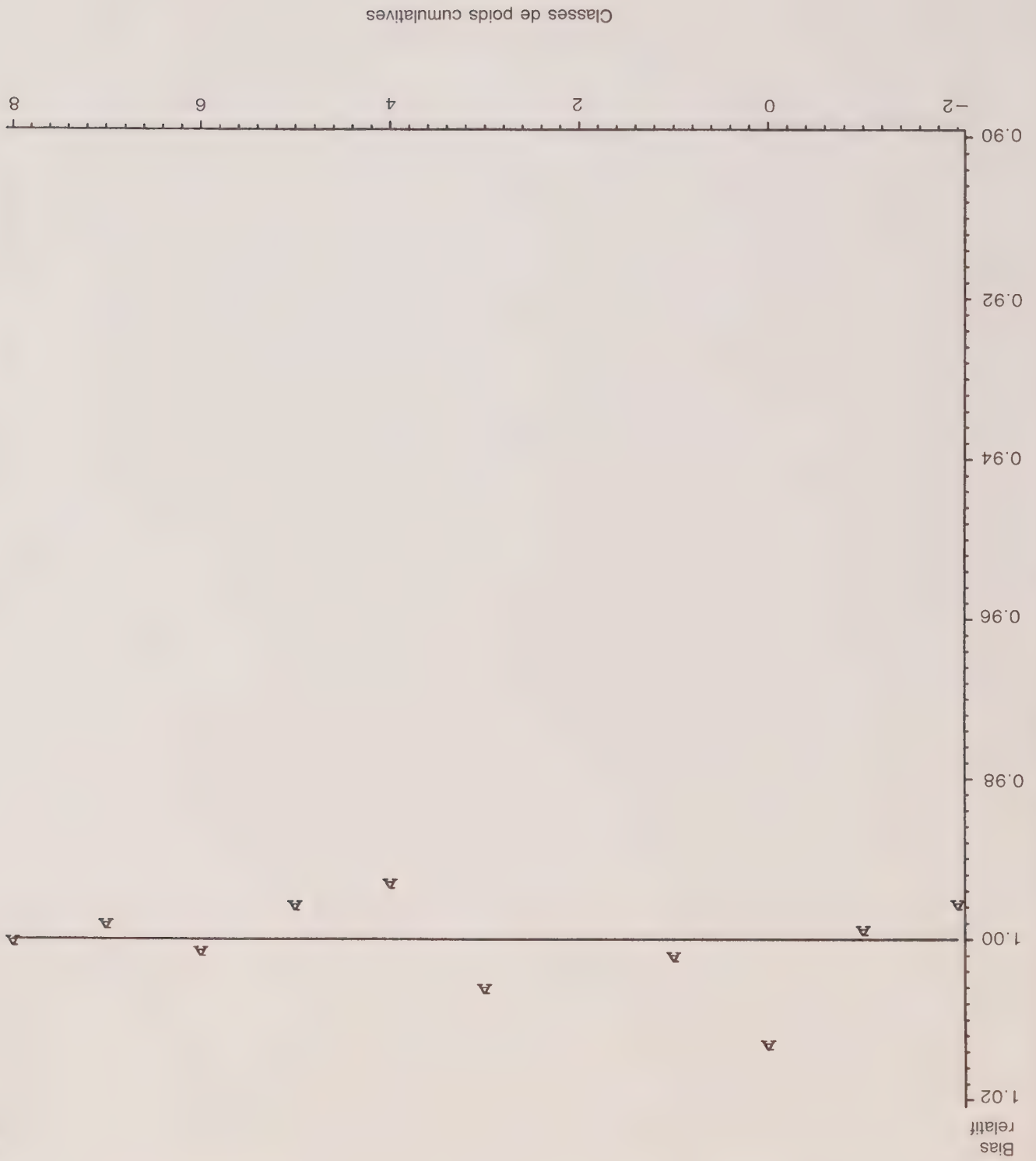


Figure 6. Biais relatif pour les estimateurs non redressés

Figure 5. Biases relatifs pour les estimateurs redressés probabilités estimées



ayant un poids de 15 ou plus sont combinées dans la classe 15+, et toutes les paires ayant un poids de – 9 ou moins sont combinées dans la classe – 10 –. Bien qu'il ait été possible, dans le cas des résultats de Rubin-Belin, d'effectuer des redressements individuels pour tenir compte des probabilités de couplage, nous avons choisi d'effectuer des redressements moyens, pour chaque classe de poids du tableau 1. (Voir Czajka et coll. 1992, pour des observations sur une décision semblable.) Notre méthode de rapproche un peu des travaux relatifs aux scores de propension (p. ex. Rosenbaum et Rubin 1983, 1985). Les techniques de mesure de la propension, bien qu'elles aient été proposées pour d'autres catégories de problèmes, pourraient aussi s'appliquer dans le présent cas.

4. POINTS SAILLANTS ET LIMITES DES RÉSULTATS DES SIMULATIONS

En raison de contraintes d'espace, nous ne présenterons que quelques résultats représentatifs des simulations effectuées. Pour plus d'information, y compris un vaste ensemble de tableaux, voir Winkler et Scheuren (1991).

Les deux mesures résultant de nos simulations auxquelles nous nous attarderons sont le biais relatif et l'erreur-type relative. Nous n'examinerons de façon détaillée que le scénario de couplage médiocre, en nous limitant au cas où R^2 se situe entre 0.40 et 0.45. Les figures 5 à 7 présentent le biais relatif pour un échantillon représentatif unique. Toutefois, un sommaire global visant les autres scénarios est présenté au tableau 2. Certaines limites des simulations sont également signalées à la fin de la présente section.

4.1 Exemple de résultats – couplage médiocre

Plutôt que d'utiliser toutes les paires, nous ne tenons compte que de celles ayant un poids de 10 ou moins. L'utilisation de ce sous-ensemble nous permet d'examiner des méthodes de redressement des données de régression pour des classes de poids comportant une proportion de non-liens vrais allant de faible à élevée. Notons que les paires éliminées (ayant un poids de 10 ou plus) correspondent seulement à des liens vrais. Les figures 5 et 6 présentent nos résultats pour les données de régression redressées et non redressées, respectivement. Les résultats obtenus avec les données non redressées sont basés sur des formules de régression classiques (p. ex. Draper et Smith 1981). Les paires ayant le poids le plus élevé. La classe de poids w correspond à toutes les paires ayant un poids entre w et 10. Nous observons ce qui suit:

- Le cumul se fait par poids d'appariement décroissants (c.-à-d. à partir des classes les plus susceptibles de contenir presque seulement des liens vrais, en allant vers les classes contenant des proportions croissantes de non-liens vrais). En particulier, pour la classe de poids

- 8, premier point de données indiqué aux figures 5 à 7, il y avait 3 non-liens et 439 liens. Une fois les données cumulées, disons, jusqu'à la classe $w = 5$, il y avait 24 non-liens, le nombre de liens, toutefois, avait grimpé à 1,121 – ce qui nous donnait une taille d'échantillon globale beaucoup plus grande, avec une réduction correspondante de l'erreur-type de régression.
- Les biais relatifs sont fournis pour le coefficient de pente original et redressé \hat{a}_1 sous forme du rapport entre le coefficient vrai (environ 2) et le coefficient calculé pour chaque classe de poids cumulative.
- Les résultats de régression redressés se fondent aussi bien sur les probabilités de concordance estimées que sur les probabilités vraies. En particulier, la figure 5 correspond aux résultats obtenus d'après les probabilités estimées (qui seraient normalement les seules disponibles en pratique). La figure 7 correspond à la situation irréaliste dans laquelle nous connaissons les probabilités vraies.

- Les erreurs quadratiques moyennes relatives (non présentes) sont obtenues par le calcul de l'EQM de chaque classe de poids cumulative. Pour chaque classe, le biais est élevé au carré, ajouté au carré des erreurs-types, puis la racine carrée est prise.

Les observations faites se déduisent assez directement de nos résultats et sont à peu près celles que nous avions prévues. Par exemple, à mesure que la taille d'échantillon augmentait, nous avons constaté que les erreurs quadratiques moyennes relatives diminuaient passablement pour les coefficients redressés. Si les coefficients de régression n'étaient pas redressés, les erreurs-types diminuaient quand même avec l'augmentation de la taille d'échantillon, mais au prix d'une augmentation inacceptable du biais. Un sujet d'inquiétude a trait à notre capacité d'estimer avec précision les probabilités de concordance, car elles ont un effet crucial sur l'exactitude des estimations des coefficients. Si nous pouvons estimer précisément les probabilités (comme c'est le cas ici), la méthode de redressement fonctionne raisonnablement bien; si nous ne pouvons pas le faire (voir plus loin), la méthode de redressement pourrait donner de piètres résultats.

4.2 Sommaire général des résultats

Nos résultats ont varié quelque peu pour les trois différentes valeurs de R^2 , se révélant meilleurs pour les R^2 les plus élevés. Ces différences conduisent, toutefois, ne modifient pas nos principales conclusions; le tableau 2, par conséquent, n'en fait pas état. Notons que dans le cas du scénario de couplage efficace, l'effort de redressement est peu bénéfique et peut même être légèrement préjudiciable. Il est en tout cas inutile, et nous ne l'avons inclus dans nos simulations que par souci d'exhaustivité. À l'autre extrême, même pour un couplage pauvre, nous avons obtenu des résultats satisfaisants, mais seulement lorsque les probabilités vraies étaient utilisées – ce qui est impossible en pratique.

Tableau 1

Nombre de liens vrais et de non-liens vrais et de non-liens vrais de lien erroné par intervalle de poids pour divers scénarios de couplage; probabilités estimées selon la méthode de Rubin-Bellin

Taux de concordance fausses

Poids	Scénario efficace			Scénario médiocre			Scénario pauvre		
	Liens			Liens			Liens		
	Vrais	NL	Est	Vrais	NL	Est	Vrais	NL	Est
15 +	9,176	0	.00	2,621	0	.00	0	.00	.00
14	111	0	.00	418	0	.00	0	.00	.00
13	91	0	.00	1,877	0	.00	0	.00	.00
12	69	0	.00	1,202	0	.00	0	.00	.00
11	59	0	.00	832	0	.00	0	.00	.00
10	69	0	.00	785	0	.00	0	.00	.00
9	42	0	.00	610	0	.00	0	.00	.00
8	36	2	.05	439	3	.00	1	.02	.00
7	30	1	.03	250	4	.00	1	.03	.00
6	14	7	.33	265	9	.03	57	.03	.03
5	28	4	.12	167	8	.05	56	.03	.03
4	6	3	.33	89	6	.06	62	.02	.05
3	12	7	.37	84	5	.06	31	.02	.11
2	8	6	.43	38	7	.16	947	.03	.19
1	7	13	.65	33	34	.51	516	.18	.25
0	7	4	.36	13	19	.61	114	.20	.28
-1	3	5	.62	7	20	.74	23	.20	.31
-2	0	11	.99	3	11	.79	23	.38	.41
-3	4	6	.60	4	19	.83	69	.82	.60
-4	4	3	.43	0	15	.99	70	.99	.70
-5	4	4	.50	0	15	.96	25	.99	.68
-6	0	5	.99	0	27	.99	85	.99	.67
-7	1	6	.86	0	40	.99		.99	.99
-8	0	8	.99	0	41	.99		.99	.99
-9	0	4	.99	0	4	.99		.99	.99
-10 -	0	22		0	22	.99		.99	.99

Notes: Dans la première colonne, le poids 10 signifie l'intervalle entre 10 et 11. Les classes de poids 15 et supérieures, ainsi que les classes de poids 9 et inférieures, ont été combinées. Les poids sont des logarithmes de rapports basés sur les probabilités de concordance estimées. NL désigne les non-liens et Prob. désigne la probabilité.

Signalons qu'il y a deux raisons pour lesquelles nous avons créé les données (x, y) utilisées dans les analyses. Premièrement, nous voulions exercer sur les données de régression un contrôle suffisant pour être en mesure de déterminer quel était l'effet de l'erreur de couplage. Il s'agissait d'un aspect important dans le cas des très vastes simulations de Monte Carlo dont il est fait état dans Winkler et Scheuren (1991). Deuxièmement, nous n'avions pas à notre disposition une paire de bases de données pour laquelle une information de couplage de grande précision était disponible et qui contenait des données quantitatives convenables.

Dans le cadre de nos simulations, dont certaines sont présentées ici, nous avons créé plus de 900 bases de données, correspondant à un grand nombre de variantes des trois

scénarios de couplage de base. Chaque base de données contenait trois paires de variables (x, y) correspondant aux trois scénarios de régression de base. Nous avons effectué une analyse de ces bases de données pour examiner la sensibilité des régressions au couplage et les redressements connexes de la méthode d'échantillonnage. Les différences de bases de données correspondant aux diverses valeurs de départ sont appelées *échantillons différents*.

Les redressements des données de régression ont été effectués séparément pour chaque classe de poids pré-sentée au tableau 1, en utilisant aussi bien les probabilités de couplage estimatives que les probabilités vraies. Dans le tableau 1, la classe de poids 10 vise les paires ayant un poids entre 10 et 11, tandis que la classe de poids 1 - 1 vise les paires ayant un poids entre 1 et 2. Toutes les paires

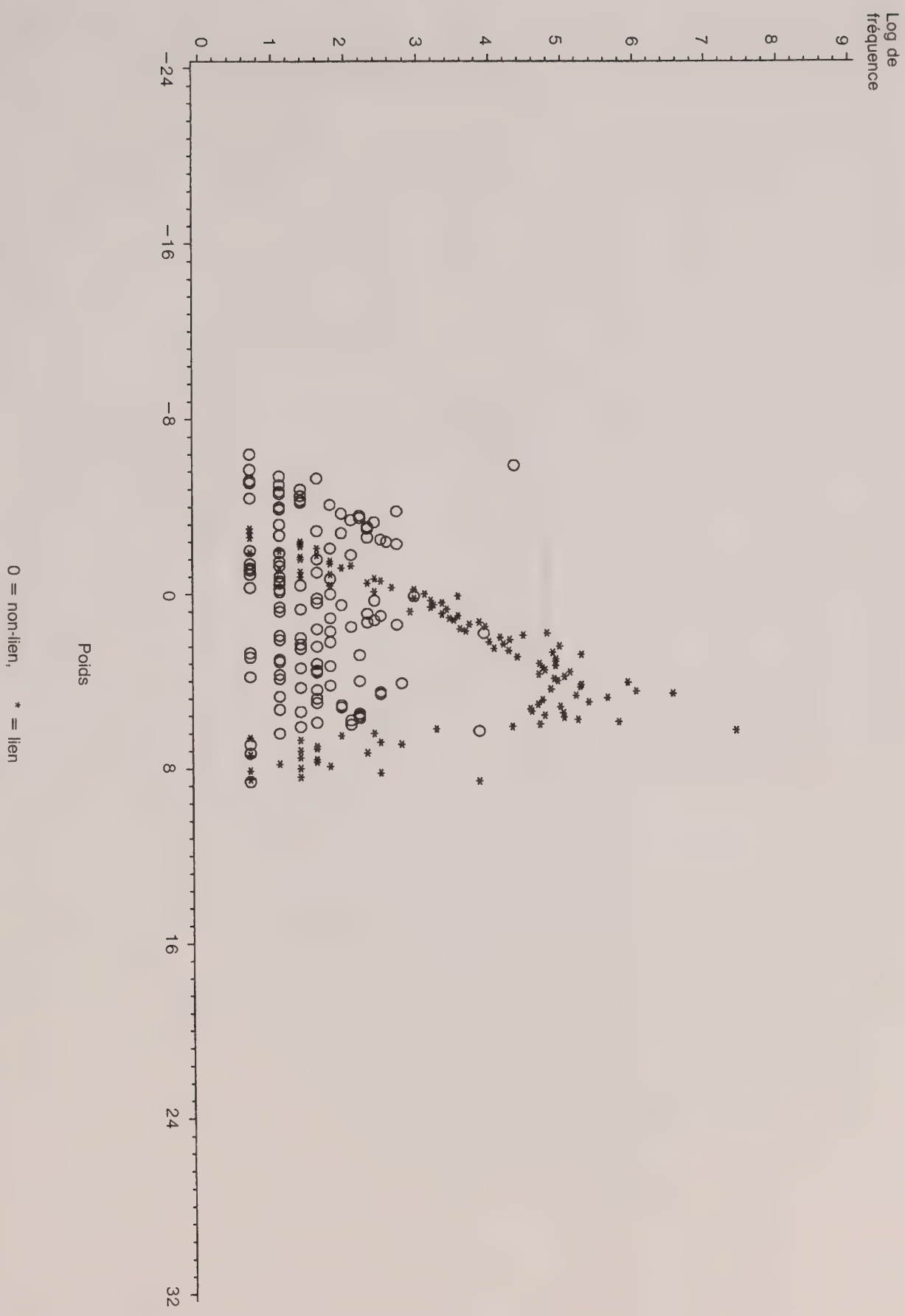


Figure 4. Logarithme de fréquence vs poids, Scénario de couplage pauvre, liens et non-liens

0 = non-lien, * = lien

Poids

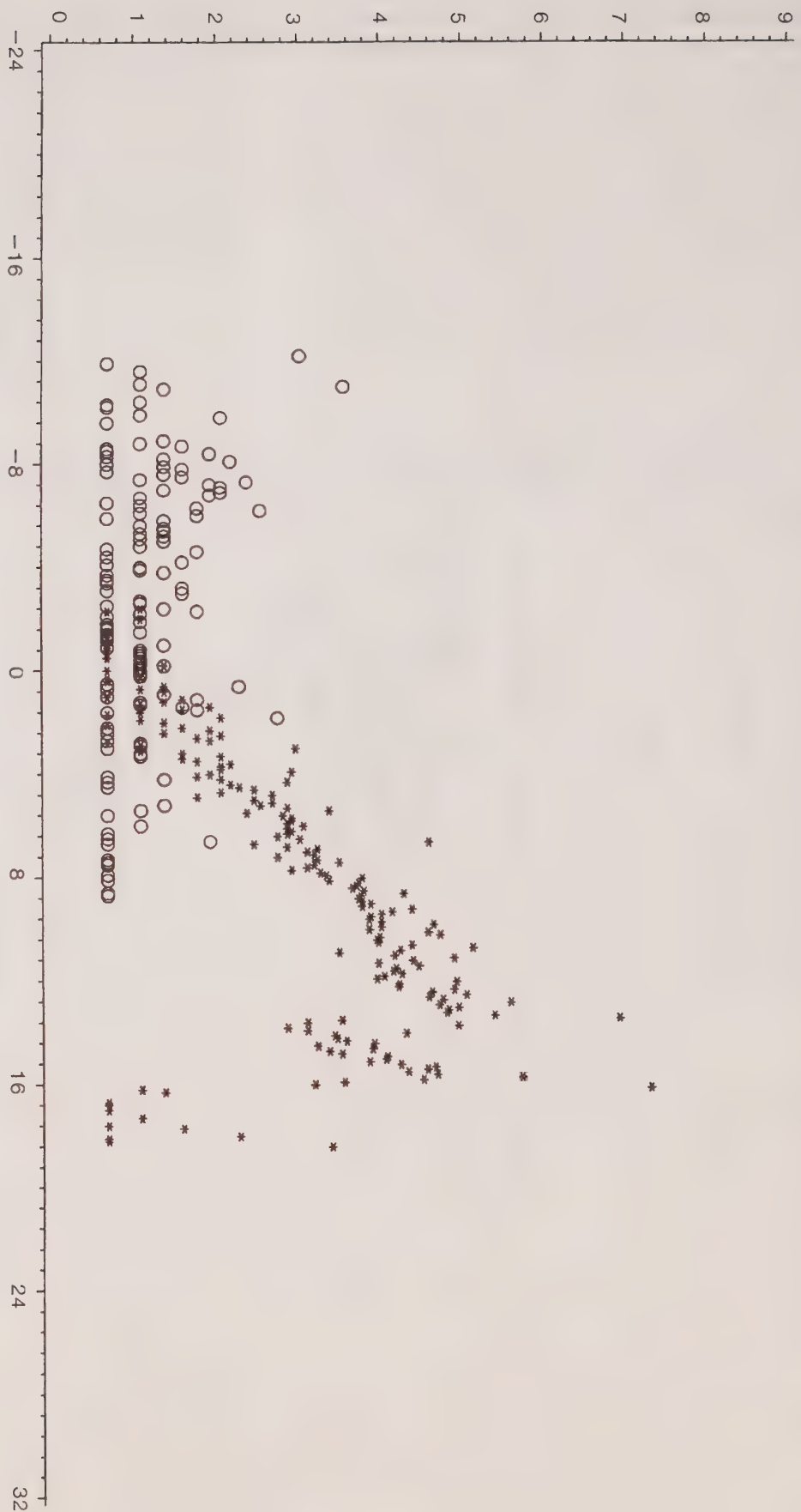


Figure 3. Logarithme de fréquence vs poids, Scénario de couplage médiocre, liens et non-liens

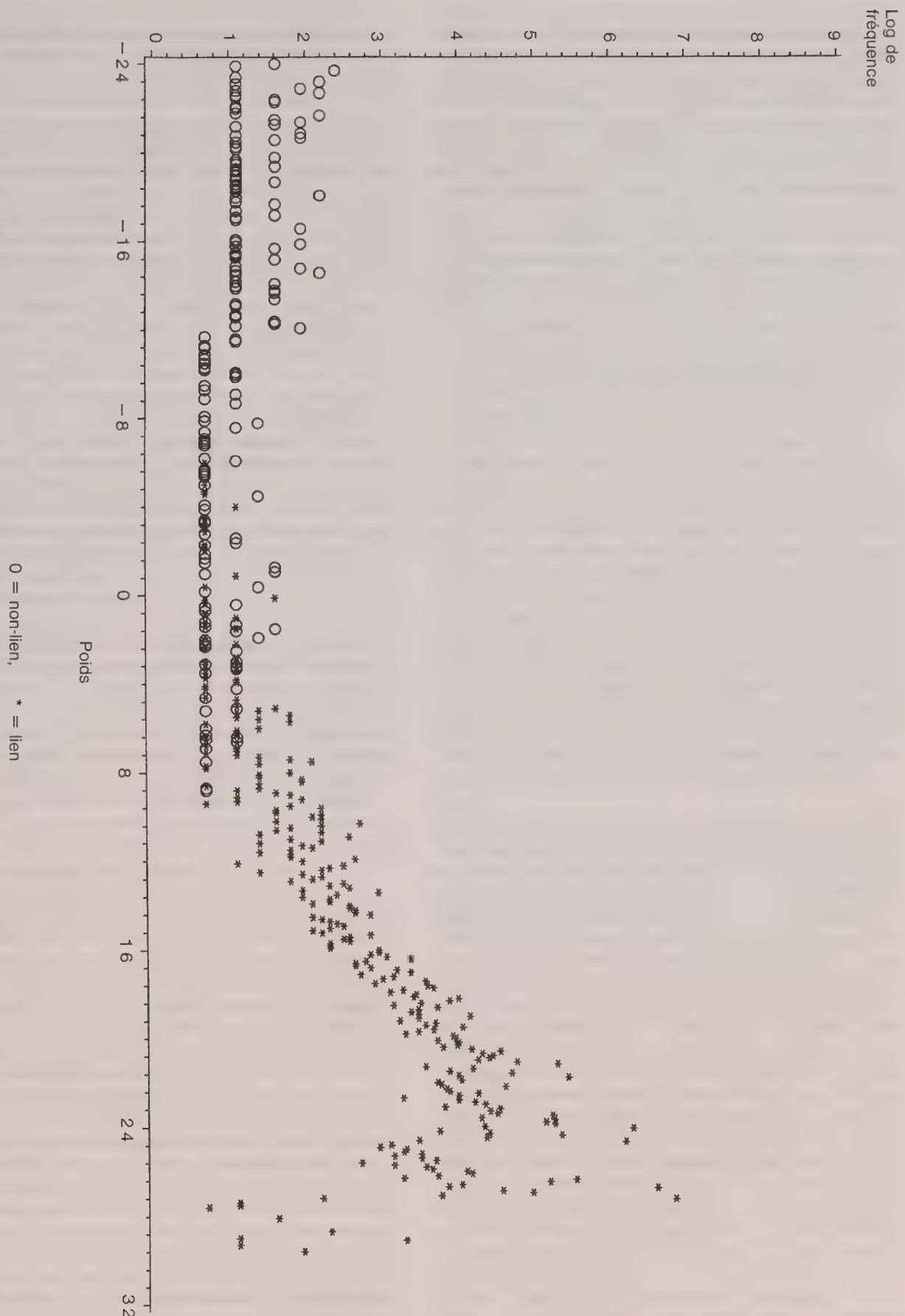


Figure 2. Logarithme de fréquence vs poids, Scénario de couplage efficace, liens et non-liens

méthode de base a consisté à prendre deux fichiers dont les concordances réelles étaient connues et à refaire le couplage selon différentes variables – ou en réalité des versions des mêmes variables avec introduction de différents degrés de distorsion, rendant de plus en plus difficile la distinction entre un lien et un non-lien. Nous avons ainsi créé un cadre comportant assez de pouvoir discriminant pour que l'algorithme de Rubin-Bellin visant l'estimation des probabilités fonctionne, mais pas assez de pouvoir discriminant pour que la zone de chevauchement des liens possibles devienne non significative.

Nous avons obtenu les résultats de base de la simulation en commençant avec une paire de fichiers de taille 10,000 comportant une information utile pour les besoins du couplage et dont les concordances réelles étaient connues. Pour les besoins des simulations, divers degrés d'erreur ont été introduits dans les variables de comparaison, différentes quantités de données ont été utilisées pour le couplage et des écarts plus grands par rapport aux probabilités de concordance optimales ont été permis.

Trois scénarios de couplage ont été examinés: (1) *efficace*, (2) *médiocre*, (3) *pauvre*. Le scénario de couplage efficace a consisté à utiliser la plupart des méthodes disponibles qui avaient servi au couplage durant le recensement américain de 1990 (p. ex. Winkler et Thibaudeau 1991). Les variables de comparaison étaient le nom, le prénom, l'initiale, le numéro de maison, le nom de rue, l'identificateur d'appartement ou d'unité, le numéro de téléphone, l'âge, l'état matrimonial, le lien avec le chef de ménage, le sexe et la race. Les probabilités de concordance utilisées dans les ratios de vraisemblance cruciaux nécessaires aux règles de décision ont été fixées à des niveaux voisins du niveau optimal.

Le scénario de couplage médiocre a consisté à utiliser le nom, le prénom, l'initiale, deux variantes de l'adresse, l'identificateur d'appartement ou d'unité, et l'âge. De légères erreurs typographiques ont été introduites de façon indépendante dans un septième des noms de famille et un cinquième des prénoms. Les probabilités de concordance ont été fixées à des niveaux s'écartant du niveau optimal, mais demeurant comparables à celles que pourrait sélectionner un spécialiste expérimenté du couplage par ordinateur.

Dans le scénario de couplage pauvre, les variables de comparaison utilisées étaient le nom, le prénom, une variante de l'adresse et l'âge. De légères erreurs typographiques ont été introduites de façon indépendante dans un cinquième des noms de famille et un tiers des prénoms. Des erreurs typographiques assez graves ont été introduites dans un quart des adresses. Les probabilités de concordance ont été fixées à des niveaux s'écartant substantiellement du niveau optimal. Nous visions ainsi à simuler le choix que pourrait faire un responsable du couplage n'ayant que peu d'expérience dans ce domaine.

Notre capacité de faire la distinction entre les liens vrais et les non-liens varie sensiblement d'un scénario à l'autre. Dans le cas du scénario efficace, nous voyons que le diagramme de dispersion des liens et des non-liens vrais se compose de deux parties presque entièrement séparées

Ainsi, nous avons transformé le scénario de couplage efficace en un scénario de couplage pauvre (figures 2 à 4) principalement en réduisant l'information de comparaison et en introduisant des erreurs typographiques dans les variables de comparaison. Même si nous avions gardé les mêmes variables de comparaison que dans le scénario de couplage efficace (figure 2), nous aurions pu obtenir un chevauchement des courbes (comme à la figure 4) simplement en faisant varier les paramètres d'appariement donnés par l'équation (2.1). Le scénario de couplage pauvre peut se produire lorsqu'on ne dispose pas d'un logiciel convenable d'analyse syntaxique des noms permettant la comparaison des noms de maison et des noms de rue correspondants. En l'absence d'une analyse syntaxique appropriée, des variables de comparaison déterminantes pour la désignation de nombreux liens vrais ne seront pas adéquatement utilisées.

Notre capacité d'estimer la probabilité d'une concordance varie beaucoup. Les valeurs vraies et estimées de ces probabilités, par classes de poids, sont présentées au tableau 1. Pour les scénarios de couplage efficace et médiocre, les probabilités estimées étaient assez voisines des valeurs vraies. Dans le cas du scénario pauvre, dans lequel la plupart des paires sont des liens possibles, les écarts sont très prononcés.

Pour chaque scénario de couplage, des données empiriques ont été créées. Chaque base de données comprenait un poids d'appariement par ordinateur, des probabilités de concordance vraies et estimées, la variable *x* indépendante pour la régression, la variable *y* dépendante, la variable *y* observée dans l'enregistrement ayant le poids d'appariement ayant le plus élevé, et la variable *y* observée dans l'enregistrement ayant le poids d'appariement venant au deuxième rang.

Les variables *x* indépendantes pour la régression ont été construites à l'aide de la procédure RANUNI du SAS, de façon à être uniformément distribuées entre 1 et 101. Pour les fins du présent article, elles ont été choisies indépendamment de toute variable de comparaison. (Bien que nous ayons examiné la situation dans laquelle les variables de régression dépendent d'une ou plusieurs variables de comparaison (Winkler et Scheuren 1991), nous ne présentons pas ces résultats ici.)

Trois scénarios de régression, correspondant à des valeurs de R^2 de plus en plus faibles, ont ensuite été examinés, c.-à-d. R^2 (1) entre 0.75 et 0.80; (2) entre 0.40 et 0.45; et (3) entre 0.20 et 0.22. Les variables dépendantes ont été créées avec des valeurs de départ indépendantes au moyen de la procédure RANNOR du SAS. Dans chaque scénario de couplage (efficace, médiocre ou pauvre), tous les appartements d'enregistrements produits par le processus de couplage et, donc, l'erreur de couplage, étaient fixes.

fichier) et les variables x (d'un autre fichier) ne représentent pas toujours la même unité.

Dans une telle situation, l'estimateur non redressé de a_1 serait biaisé; toutefois, en vertu d'hypothèses comme celle qui consiste à supposer x et y indépendants en cas d'erreur de couplage, il peut être démontré que, si nous connaissons le taux d'erreurs de couplage h , nous pouvons obtenir un estimateur redressé non biaisé par une simple correction de l'estimateur ordinaire, en le multipliant par $(1/(1-h))$. Intuitivement, on peut croire que les paires faisant l'objet d'une erreur de couplage entraînent une sous-estimation de la véritable corrélation (positive ou négative) entre x et y . Le coefficient redressé compense pour cette sous-évaluation. Une fois connu le coefficient de pente redressé a_1 , l'ordonnée à l'origine appropriée peut être obtenue de l'expression habituelle $a_0 = \bar{y} - a_1\bar{x}$, où \bar{a}_1 est le coefficient redressé.

Des méthodes d'estimation des erreurs-types de régression peuvent aussi être élaborées pour les cas où il existe des erreurs de couplage. Plutôt que de continuer d'examiner ce cas spécial, toutefois, nous allons examiner comment l'idée de faire un redressement multiplicatif peut être généralisée. Considérons

$$Y = X\beta + \epsilon, \quad (3.2)$$

le modèle de régression unidimensionnel ordinaire, pour lequel tous les termes d'erreur ont comme moyenne zéro et sont indépendants avec variance constante σ^2 . Si nous travaillions avec une base de données de taille n , nous ferions une régression de Y en X de la façon normale. Mais, puisque chaque enregistré est apparié à deux autres, nous disposons globalement de $2n$ paires. Nous souhaitons utiliser (X_i, Y_i) , mais nous utilisons plutôt (X_i, Z_i) . Z_i peut Y_i , mais peut prendre aussi une autre valeur, X_j , en raison de l'erreur de couplage.

$$\text{Pour } i = 1, \dots, n,$$

$$Z_i = \begin{cases} Y_i & \text{avec probabilité } p_i \\ X_j & \text{avec probabilité } q_j \text{ pour } j \neq i, \end{cases} \quad (3.3)$$

$$p_i + \sum_j q_j = 1.$$

La probabilité p_i peut être zéro ou un. Nous définissons $h_i = 1 - p_i$ et divisons l'ensemble de paires en n classes mutuellement exclusives. Les classes sont établies d'après les enregistrements d'un des fichiers. Chaque classe comprend la variable x indépendante X_j , la valeur vraie de la variable y dépendante, les valeurs des variables y des enregistrements du deuxième fichier avec lesquels l'enregistrement du premier fichier contenant X_i forme une paire, et les probabilités (ou poids) du couplage par ordinateur. Les liens, les non-liens et les liens possibles sont inclus. Dans une hypothèse de couplage binivoque, pour chaque $i = 1, \dots, n$, il existe au plus un j tel que $q_j > 0$. Nous posons aussi $\phi, \text{ défini par } \phi(i) = j$.

L'idée intuitive de notre approche (et de celle de Neter et coll.) est que nous pouvons, en vertu des hypothèses du modèle, exprimer chaque paire de points de données observées (X, Z) en termes des valeurs vraies (X, Y) et d'un terme de biais (X, b) . Toutes les équations nécessaires aux techniques de régression habituelles peuvent alors être obtenues. Nos formules de calcul sont beaucoup plus complexes que celles de Neter et coll. car leur hypothèse contraignante (3) a permis une importante simplification de leurs formules de calcul. En particulier, en vertu de leurs hypothèses, Neter et coll. ont prouvé que la moyenne et la variance des valeurs Z observées étaient toutes deux nécessairement égales à la moyenne et à la variance des valeurs Y vraies.

En vertu du modèle considéré ici, nous observons (voir l'annexe) que

$$E(Z) = (1/n) \sum_i E(Z|i) = (1/n) \sum_i (X_i p_i + \sum_j X_j q_{ij})$$

$$= (1/n) \sum_i X_i + (1/n) \sum_i [X_i(-h_i) + X_{\phi(i)} h_i] = \bar{Y} + B. \quad (3.4)$$

Puisque chaque $X_i, i = 1, \dots, n$, peut être apparié avec soit Y_i , soit $X_{\phi(i)}$, la deuxième égalité dans (3.4) représente $2n$ points. De la même façon, nous pouvons représenter σ_{xy}^2 en fonction de σ_{xy}^2 et d'un terme de biais B_{xy} , et σ_y^2 en fonction de σ_y^2 et d'un terme de biais B_y . Nous ne supposons ni que les termes de biais ont une espérance zéro ni qu'ils ne sont pas corrélés avec les données observées.

Avec les différentes représentations, nous pouvons redresser les coefficients de régression β_x et leurs erreurs-types associées et les ramener aux valeurs vraies β_{yx} et à leurs erreurs-types associées. Notre hypothèse d'un couplage binivoque (qui n'est pas nécessaire à la théorie générale) est posée en vue de faciliter les calculs et de réduire le nombre d'enregistrements ainsi que la quantité d'information devant être prise en considération au cours du processus de couplage.

Pour les besoins des redressements, nous faisons deux hypothèses cruciales. La première est que, pour $i = 1, \dots, n$, nous pouvons estimer avec exactitude les probabilités vraies d'une concordance p_i . Voir l'annexe pour une description de la méthode de Rubin et Belin (1991). La deuxième est que, pour chaque $i = 1, \dots, n$, la valeur vraie X_i associée à la variable indépendante X_i est celle de la paire ayant le poids d'appariement le plus élevé, tandis que la valeur faussée $X_{\phi(i)}$ est celle de la paire dont le poids d'appariement vient au deuxième rang. (D'après les simulations effectuées, il semble qu'au moins la première de ces deux hypothèses ait un grand rôle lorsqu'une portion importante des paires sont des liens possibles.)

3.2 Application simulée

A l'aide des méthodes qui viennent d'être décrites, nous avons tenté une simulation avec des données réelles. Notre

de rechange fiables à une vérification complète? Nous croyons que oui, et cette conviction motive la perspective que nous adoptons à la section 3, où nous examinons les erreurs de couplage dans un contexte d'analyse de régression. D'autres approches, toutefois, pourraient être nécessaires dans des cadres d'analyse différents.

3. RÉGRESSION PORTANT SUR DES DONNÉES COUPLÉES

Pour les besoins de notre examen de la régression, nous faisons l'hypothèse que le responsable du couplage a aidé l'analyste en lui fournissant un fichier de données combiné comprenant des paires d'enregistrements – un de chacun des fichiers de départ – ainsi que la probabilité de concordance et l'état de concordance de chaque paire. Les liens, les non-liens et les liens possibles sont tous inclus et bien identifiés. Il semble évident qu'il faille garder les liens probables et les liens possibles, mais il est moins évident que les non-liens probables soient utiles. Toutefois, comme l'a signalé Newcombe, l'information sur les non-liens probables est nécessaire au calcul des biais. Nous faisons l'hypothèse qu'il sera suffisant de garder au maximum deux ou trois paires comprenant un enregistrement du fichier B pour chaque enregistrement du fichier A. Les deux ou trois paires ayant les poids d'appariement les plus élevés seraient conservées.

Plus précisément, nous supposons que le fichier des liens a été élargi de telle façon que chaque enregistrement du plus petit des deux fichiers a été apparié avec, disons, les deux enregistrements du fichier le plus gros ayant les poids d'appariement les plus élevés. Puisque $n \leq m$, nous conservons $2n$ des $n \times m$ paires possibles. Pour chaque enregistrement, nous conservons les indicateurs de concordance et les probabilités associées aux enregistrements avec lesquels il forme une paire. Dans certains cas, nous aurons des combinaisons (lien, non-lien) ou des combinaisons (non-lien, non-lien). À des fins de simplicité, nous ne traiterons pas des situations où il peut exister plus d'un lien vrai; par conséquent, les combinaisons (lien, lien) sont par définition écartées.

Comme on peut s'en rendre compte, une telle structure de données se prête à une analyse par différentes méthodes. Par exemple, nous pouvons décomposer le fichier en trois parties – liens désignés, non-liens et liens possibles. L'analyse, quelle qu'elle soit, pourrait être appliquée séparément à chaque groupe ou à des sous-ensembles de ces groupes. Dans la méthode appliquée ici, nous utiliserons les non-liens pour redresser l'ensemble des liens potentiels et, ainsi, obtenir une perspective additionnelle pouvant se traduire par une erreur quadratique moyenne (EQM) moindre que celle de statistiques calculées uniquement à partir des données sur les liens.

Pour les analyses statistiques, si nous devons utiliser seulement les données relatives aux paires d'enregistrements qui constituent presque assurément des liens, nous nous privons peut-être d'une bonne part d'information additionnelle provenant de l'ensemble des paires formant

des liens possibles qui, en tant que sous-ensemble, pourrait contenir autant de liens vrais que l'ensemble des paires désignées comme des liens. En outre, nous pourrions introduire des biais importants dans les résultats, car certains sous-ensembles des liens vrais susceptibles de nous intéresser pourraient résider principalement dans l'ensemble des liens possibles. Par exemple, si nous analysons des aspects liés à l'action positive et aux revenus, certains enregistrements (par exemple ceux ayant trait aux personnes à faible revenu) pourraient être plus difficiles à coupler par l'information sur le nom et l'adresse et donc se trouver en forte concentration dans l'ensemble des liens possibles.

3.1 Fondement théorique

Neter, Maynes et Ramamanathan (1965) ont reconnu que les erreurs introduites au cours du processus de couplage pouvaient avoir un effet négatif sur les analyses ayant comme base les fichiers couplés résultants. Pour montrer comment les idées de Neter et coll. ont inspiré l'approche adoptée dans le présent article, nous fournissons des détails additionnels sur leur modèle. Neter et coll. ont supposé que l'ensemble des enregistrements d'un fichier (1) pouvaient être appariés, (2) avaient toujours la même probabilité p d'être appariés correctement et (3) avaient la même probabilité q d'être appariés incorrectement à n importe quel enregistrement restant du deuxième fichier (c.-à-d. $p + (N - 1)q = 1$ où N est la taille du fichier). Ils ont généralisé leurs résultats de base en supposant que les ensembles de paires formés à partir des deux fichiers pouvaient être décomposés en classes pour lesquelles (1), (2) et (3) demeuraient valables.

Notre méthode s'inspire de celle de Neter et coll. parce que nous croyons que leur approche est pertinente. Nous souscrivons à leur conclusion selon laquelle un niveau modéré d'erreur de couplage peut introduire un biais élevé dans les coefficients de régression. Nous ne croyons pas, toutefois, que la condition (3) – qui a constitué leur principal moyen de simplification des formules de calcul – sera jamais respectée en pratique. Si le couplage se fonde sur des identificateurs uniques comme les numéros de sécurité sociale, qui sont sujets aux erreurs typographiques, il est improbable qu'un enregistrement comportant une erreur typographique aura la même probabilité d'être couplé incorrectement à tous les enregistrements restants du deuxième fichier. Si les variables de comparaison sont le nom et l'adresse (qui souffrent souvent d'un nombre beaucoup plus grand d'erreurs typographiques), le respect de la condition (3) est encore plus improbable.

Pour bien montrer comment nos travaux sont un prolongement et une généralisation des résultats de Neter et coll. nous allons examiner un cas spécial. Supposons que nous utilisions les moindres carrés ordinaires dans le cadre d'une régression simple de la forme

$$y = a_0 + a_1x + \epsilon. \tag{3.1}$$

Supposons, en outre, que des erreurs de couplage se soient produites, c'est à dire que les variables y (d'un

les examiner manuellement, pour tenter de repérer avec exactitude les liens vrais. Quant aux paires désignées par erreur comme des non-liens, on peut aussi effectuer dans leur cas une vérification supplémentaire (peut-être elle aussi manuelle). Dans les deux cas, une telle démarche est coûteuse, longue et sujette aux erreurs.

Fait peu surprenant, la recherche dans le domaine du couplage d'enregistrements a surtout porté, depuis les travaux initiaux de Newcombe, sur des moyens de réduire les étapes de vérification manuelle destinées à l'examen des liens possibles. De vastes progrès ont été accomplis en ce qui a trait à l'amélioration des règles de couplage, grâce à une meilleure utilisation de l'information contenue dans les paires d'enregistrements, et à l'estimation des taux d'erreur par l'entremise de modèles probabilistes.

Diverses méthodes ont été mises à contribution pour améliorer les règles de couplage d'enregistrements. Pour traiter les erreurs typographiques légères, comme le fait d'écrire "Smith" au lieu de "Smith", Winkler et Thibaudau (1991) ont proposé une extension du comparateur de chaînes de Jaro (1989). Newcombe et coll. (1989), pour leur part, ont élaboré des méthodes pour la création et l'utilisation de tableaux de concordance partielle. Pour certaines catégories de tableaux de fichiers, Winkler et Thibaudau (1991) (voir aussi Winkler 1992; Jaro 1989) ont élaboré des algorithmes espérance-maximisation, ainsi que des techniques de modélisation fondées sur une information *a priori* qui donnait automatiquement les paramètres optimaux de (2.1) en vue de leur utilisation dans les règles de décision (2.2).

Rubin et Belin (1991) ont introduit une méthode permettant l'estimation des taux d'erreur lorsque ceux-ci ne peuvent pas être estimés de façon fiable par les méthodes habituelles (Belin 1991, p. 19-20). En utilisant un modèle selon lequel les courbes "poids vs. logarithme de fréquence" produites par le processus de couplage pouvaient être exprimées sous forme d'une combinaison de deux courbes (liens et non-liens), Rubin et Belin ont estimé les courbes qui, à leur tour, donnaient des estimations des taux d'erreur. Pour appliquer leur méthode, Rubin et Belin avaient besoin d'un échantillon d'apprentissage donnant une estimation *a priori* de la forme des deux courbes.

Bien que de nombreux problèmes de couplage surviennent rétrospectivement, souvent dans le cadre d'études épidémiologiques, il est parfois arrivé que des responsables du couplage soient en mesure d'indiquer quelle information était nécessaire dans les deux ensembles de données d'après les besoins connus de l'analyse. En exigeant une meilleure information de couplage, comme cela s'est fait dans l'enquête post-dénombrément du recensement de 1990 (voir, p. ex. Winkler et Thibaudau 1991), on est parvenu à réduire au minimum les ensembles de liens possibles.

Malgré ces progrès, il est possible que le responsable du couplage et l'analyste doivent finalement recourir à une vérification manuelle. Même aujourd'hui, les coûts additionnels en temps, en argent et en erreurs résiduelles cachées peuvent encore être imposants. Y a-t-il des solutions

décision produisant des résultats (presque) aussi bons que ceux donnés par la règle (2.2) quand les paramètres vrais sont utilisés.

La méthode de Fellegi-Sunter est essentiellement un prolongement direct de la théorie classique des tests d'hypothèses au couplage d'enregistrements. Pour décrire le modèle plus en détail, supposons que nous ayons deux fichiers de tailles n et m avec – sans perte de généralité – $n \leq m$. En vertu du processus de couplage, une comparaison pourrait être faite pour l'ensemble des $n \times m$ paires possibles d'enregistrements (un élément de la paire provenant de chacun des fichiers). Une décision est alors prise quant à savoir si oui ou non chaque paire de comparaison représente la même unité ou si l'on dispose d'éléments suffisants pour déterminer s'il y a un lien ou non.

Schématiquement, il est d'usage d'organiser les $n \times m$ paires du tableau selon une mesure de la probabilité que la paire représente des enregistrements relatifs à la même unité. À la figure 1, par exemple, nous avons tracé deux courbes. La courbe du côté droit est une distribution hypothétique des n liens vrais par rapport au "poids d'appariement" (calculé d'après (2.1), mais en logarithmes naturels). La courbe du côté gauche est la distribution des $n \times (m - 1)$ paires restantes – les non-liens vrais – en fonction des poids d'appariement, également en logarithmes. En général, comme le montre la figure 1, les distributions des liens et des non-liens se chevauchent. Aux extrêmes, le recoupement n'a pas de conséquence sur la décision de couplage à prendre; toutefois, il existe une région centrale de liens possibles, disons entre " L ", " U ", et " L ", où il serait difficile, d'après la figure 1 seulement, de décider avec un quelconque degré d'exactitude s'il s'agit de liens ou de non-liens.

Le modèle de Fellegi-Sunter est valide pour n'importe quel ensemble de paires examiné. Toutefois, pour faciliter les calculs, nous pouvons faire porter notre analyse non pas sur la totalité des paires possibles dans $A \times B$, mais uniquement sur un sous-ensemble de paires, pour lesquelles les enregistrements des deux fichiers concordent au niveau de données clés, ou de "groupeage", considérées comme hautement fiables. Comme exemples de *critères de groupeage logiques*, mentionnons un identificateur géographique comme le code postal (p. ex. ZIP), un identificateur de nom de famille comme un code Soundex ou NYSIS (voir, p. ex. Newcombe 1988, p. 182-184). Signalons, en passant, que le modèle de Fellegi-Sunter ne présuppose pas (comme nous le faisons à la figure 1) que parmi les $n \times m$ paires, il y aura n liens, mais plutôt, s'il n'y a pas d'enregistrements en double dans A ou B , qu'il y aura au maximum n liens.

2.2 Traitement des liens possibles

Même lorsqu'un système de couplage par ordinateur utilise la règle de décision de Fellegi-Sunter pour désigner de façon quasi certaine des paires comme des *liens vrais* ou des *non-liens vrais*, il peut rester un imposant sous-ensemble de paires qui ne représentent que des liens possibles. L'une des façons de traiter ces liens possibles consiste à

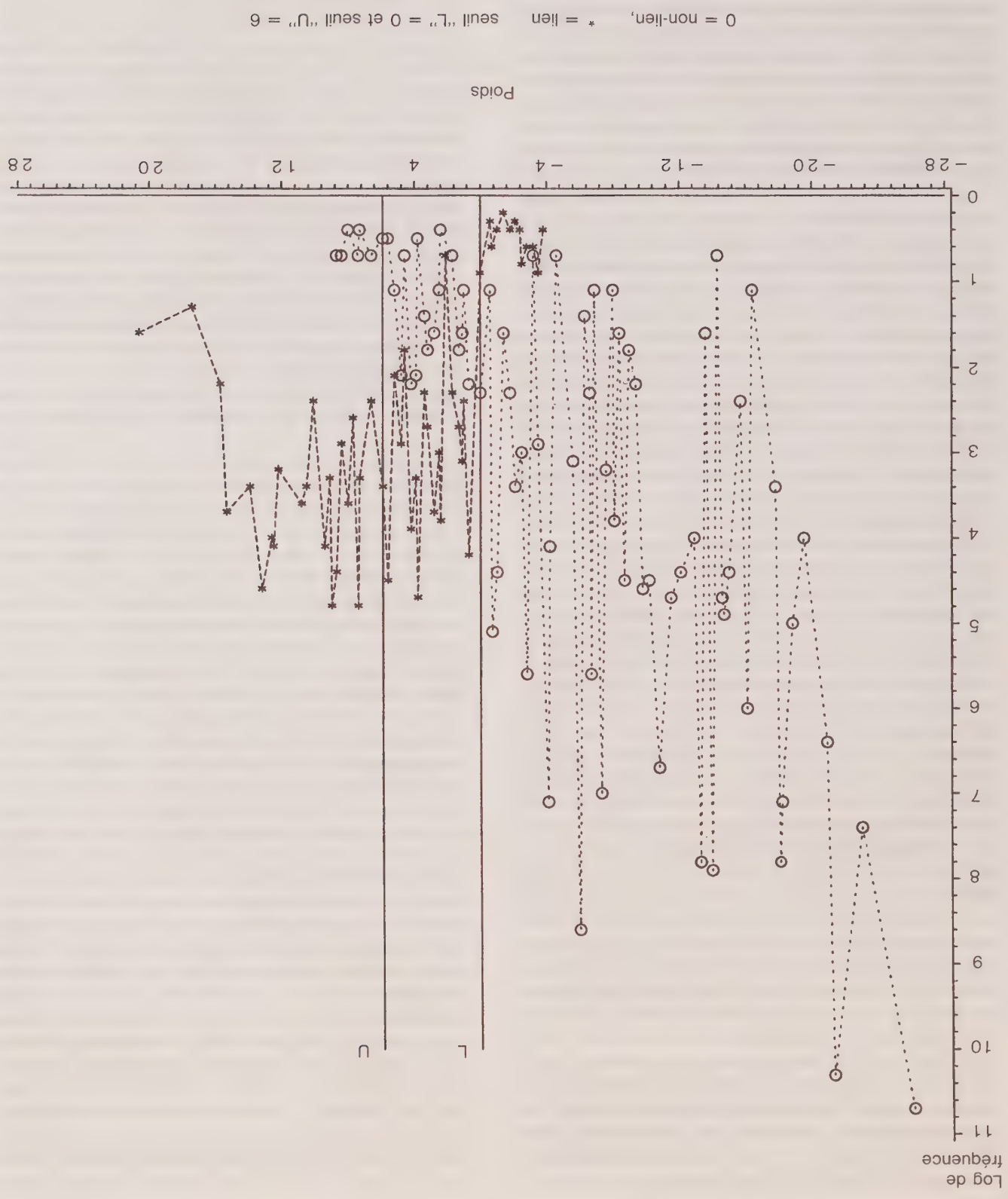


Figure 1. Logarithme de fréquence vs poids, Liens et non-liens

importante d'enregistrements non apparées qui auraient dû l'être, les analyses statistiques peuvent être satisfaisamment compromises pour que les techniques statistiques habituelles donnent des résultats trompeurs. Nous traitons essentiellement, dans le présent article, de l'impact qu'exercent les liens erronés sur les résultats des analyses. L'impact des problèmes causés par la présence de non-liens erronés (un type d'échantillonnage implicite qui peut entraîner des biais de sélection) est brièvement analysé dans la dernière section.

2.1 Modèle de couplage d'enregistrements de Fellegi-Sunter

Le processus de couplage d'enregistrements tente de classer les paires d'un espace produit $A \times B$ provenant de deux fichiers A et B en deux ensembles: M , l'ensemble des liens vrais, et U , l'ensemble des non-liens vrais. Enon-gant de façon rigoureuse des concepts introduits par Newcombe (p. ex. Newcombe et coll. 1959), Fellegi et Sunter (1969) ont examiné des rapports de probabilités de la forme:

$$R = Pr(\gamma \in T | M) / Pr(\gamma \in T | U), \tag{2.1}$$

où γ est un profil de concordance arbitraire appartenant à l'espace de comparaison T . Par exemple, T pourrait être formé de huit profils représentant une concordance simple ou non pour le nom, le prénom et l'âge. Autre possibilité, chaque $\gamma \in T$ pourrait en outre tenir compte de la fréquence relative à laquelle des noms particuliers, comme Smith ou Zabrinsky, surviennent. Les zones des enregistrements qui sont comparées (nom, prénom, âge) sont appelées *variables de comparaison*.

La règle de décision est donnée par:

Si $R > Limite\ sup.$, classer la paire comme un lien.

Si $Limite\ inf. \leq R \leq Limite\ sup.$, classer la

paire comme un lien possible et la garder pour la soumettre à une vérification manuelle. (2.2)

Si $R < Limite\ inf.$, classer la paire comme un non-lien.

Fellegi et Sunter (1969) ont montré que la règle de décision est optimale, c'est-à-dire que pour toute paire de bornes fixes de R , la région milieu est minimisée pour l'ensemble des règles de décision relatives au même espace de comparaison T . Les seuils *Limite supérieure* et *Limite inférieure* sont déterminés d'après les bornes fixes pour l'erreur. Nous appelons le rapport R ou toute transformation monotone croissante de ce dernier (donnée, par exemple, par un logarithme) un *poids d'appariement* ou un *poids de concordance totale*. Dans les applications réelles, l'optimalité de la règle de décision (2.2) est fortement liée à l'exactitude des estimations des probabilités données en (2.1). Ces probabilités sont appelées *paramètres d'appariement*. Les paramètres estimés sont (quasi) *optimaux* s'ils donnent des règles de

qui survient **après** que l'état de concordance a été déterminé. Nous supposons la situation type où le responsable du couplage effectue son travail séparément de l'analyste. Nous supposons aussi que l'analyste (ou l'utilisateur) pourrait vouloir appliquer des techniques statistiques classiques – régression, tableaux de contingence, tables de survie, etc. – au fichier résultant du couplage. Une question essentielle que l'on souhaite alors explorer est la suivante: "Que peut faire le responsable du couplage pour aider l'analyste?". La question suivante en découle: "Que devrait savoir l'analyste au sujet du couplage et comment cette information devrait-elle être utilisée?".

À notre avis, il est important de définir théoriquement les étapes du couplage et de l'analyse dans le cadre d'un système statistique unique et de concevoir les stratégies appropriées en conséquence. De toute évidence, la qualité du travail de couplage peut avoir un impact direct sur n'importe quelle analyse effectuée. Toutefois, il est rare que des mesures directes de cet impact soient fournies (p. ex. Scheuren et Oh 1975). Rubin (1990) a signalé le besoin de formuler des énoncés inférentiels destinés à résumer l'information que révèlent les données analysées. Les idées de Rubin ont été exprimées dans le contexte du recours à des techniques de préparation des données comme l'édition et l'imputation, dans des situations où, fréquemment, la non-réponse peut rendre inopérantes les méthodes statistiques courantes incluses dans les logiciels existants. Nous croyons que les idées de Rubin s'appliquent, au moins à un degré égal, au couplage d'enregistrements. La présente analyse est divisée en quatre sections. D'abord, nous faisons un rappel sur certains aspects du couplage, car toute réponse – même partielle – dépendra des fichiers à coupler et de l'utilisation faite des données couplées. Dans la section suivante, nous présentons notre cadre méthodologique en centrant notre attention, comme il a déjà été indiqué, uniquement sur l'analyse de régression. À la section 4, nous présentons quelques résultats de simulations exploratoires. Ces simulations visent à aider le lecteur à soupeser les idées que nous soumettons et à percevoir la nature de certaines difficultés. Dans la dernière section, nous présentons des conclusions préliminaires et nous suggérons des voies de recherche future. Une brève annexe, offrant plus de détails sur des aspects théoriques, est également incluse.

2. FONDEMENTS DU COUPLAGE D'ENREGLISTREMENTS

Lorsque deux fichiers ou plus sont couplés, il se peut qu'un enregistrement particulier d'un fichier ne soit pas apparié au bon enregistrement dans l'autre fichier. Si l'on ne dispose pas d'un identificateur unique pour les enregistrements correspondants des deux fichiers – ou encore si un tel identificateur souffre d'inexactitudes – le processus de couplage est sujet aux erreurs. Si la base de données couplée qui en résulte contient une part importante d'information provenant de paires d'enregistrements qui ont été apparées de façon erronée, ou encore une part

Analyse de régression de fichiers de données couplés par ordinateur

FRITZ SCHEUREN et WILLIAM E. WINKLER¹

RÉSUMÉ

Le présent article s'intéresse à la façon de traiter les erreurs de couplage d'enregistrements lorsqu'on effectue une analyse de régression. Des travaux récents de Rubin et Belin (1991) et de Winkler et Thibaudau (1991) fournissent la théorie, les algorithmes de calcul et le logiciel nécessaires à l'estimation des probabilités de concordance. Ces progrès nous permettent de mettre à jour les travaux de Neter, Maynes et Ramamanathan (1965). Des méthodes de redressement sont présentées, et certaines simulations fructueuses sont décrites. Nos résultats sont préliminaires et visent en grande partie à susciter d'autres travaux.

MOTS CLÉS: Couplage d'enregistrements; erreur de couplage; analyse de régression.

1. INTRODUCTION

L'information contenue dans deux bases de données informatiues peut être combinée à des fins d'analyse et de prise de décisions. Par exemple, un épidémiologiste pourrait être intéressé à évaluer l'effet d'un nouveau traitement contre le cancer en couplant l'information provenant d'un ensemble d'études de cas médicales et les données d'un registre de décès, afin de disposer de renseignements sur les causes et les dates des décès (p. ex. Beebe 1985). Un économiste désirant évaluer l'efficacité de décisions en matière de politique énergétique pourrait coupler une base de données contenant de l'information sur les combustibles et les matières premières pour un ensemble d'entreprises, et une base de données contenant les quantités et les types de biens produits par ces entreprises (p. ex. Winkler 1985). Si des identificateurs uniques, par exemple des numéros de sécurité sociale ou des numéros d'identification d'employeur ayant été vérifiés, sont disponibles, le couplage des sources de données peut être direct et permettre, sans autres détours, le recours aux méthodes habituelles d'analyse statistique.

Si l'on ne dispose pas d'identificateurs uniques (p. ex. Jabine et Scheuren 1986), le couplage doit être effectué d'après le nom, l'adresse, l'âge ou un autre élément descriptif de l'entreprise ou de la personne. Même en l'absence de variantes d'orthographe ou d'erreurs typographiques, il se peut que le nom, par exemple "Smith" ou "Robert", ne permette pas à lui seul d'identifier une unité. En outre, l'utilisation d'adresses entraîne souvent des erreurs attribables aux différences de formats, puisque les logiciels d'analyse syntaxique ou de normalisation existants ne permettent pas de s'assurer complètement qu'un numéro de maison ou un nom de rue, par exemple, est bien comparé à un autre numéro de maison ou un autre nom de rue. Les adresses d'une unité qu'on veut coupler peuvent également varier, soit parce que l'une d'elles est erronée, soit en raison d'un déménagement.

Depuis quelques années, beaucoup de nouveaux travaux ont été publiés en Amérique du Nord dans le domaine des techniques de couplage d'enregistrements (p. ex. Jaro 1989; et Newcombe, Fair et Lalonde 1992). Certains de ces résultats sont le produit d'une série de conférences amorcée au milieu des années 1980 (p. ex. Kilss et Alvey 1985; Howe et Spasoff 1986; Coombs et Singh 1987; Carpenter et Fair 1989); les efforts visant à étudier le sous-dénombrement dans le recensement décennal de 1990 ont constitué un autre important facteur de stimulation aux États-Unis (p. ex. Winkler et Thibaudau 1991). Le nouvel ouvrage de Newcombe (1988) a également joué un important rôle de déclencheur. Enfin, les efforts provenant d'autres sources ont aussi été considérables (p. ex. Copas et Hilton 1990). Ce qui étonne dans l'ensemble de ces travaux récents, c'est le degré de maturité atteint dans l'élaboration des principaux fondements théoriques des méthodes de couplage informatisées. Des applications pratiques bien structurées remontent au moins jusqu'aux années 1950, notamment aux travaux de Newcombe et de ses collaborateurs (p. ex. Newcombe et coll. 1959). Environ dix ans plus tard, un fondement théorique solide pour ces idées de base a été établi dans les articles de Tepping (1968) et, en particulier, de Fellegi et Sunter (1969).

Si l'intérêt à l'égard du couplage d'enregistrements s'est ainsi maintenu, c'est en partie parce que la révolution informatique a permis d'utiliser des techniques de plus en plus efficaces. La prolifération des fichiers accessibles sous forme électronique a également élargi la gamme des applications possibles. Un autre facteur d'intérêt est né du besoin d'établir des ponts entre le domaine relativement étroit (même obscur) du couplage par ordinateur et le reste du domaine des statistiques (p. ex. Scheuren 1985). Le présent article appartient à cette dernière catégorie, et vise à examiner les aspects particuliers des analyses de régression portant sur des ensembles de données couplés.

De façon générale, nous n'examinerons pas ici les techniques de couplage. Nous nous intéresserons plutôt à ce

- BIBLIOGRAPHIE**
- BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- DeGROOT, M.H. (1986). *Probabilité et Statistiques*, 2^{ème} Edition. Reading, MA: Addison-Wesley.
- FELLEGI, I.P., et SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 40, 1183-1210.
- GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 2, 215-231.
- HABERMAN, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press.
- HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.
- THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.
- WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Bureau of The Census Fifth Annual Research Conference*, 145-155.
- WINKLER, W.E. (1988). Using The E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

à indépendance conditionnelle donne d'aussi bons résultats que le modèle enrichi. Ce dernier, toutefois, est préféré parce qu'il repose sur une base théorique sans faille.

7. SUGGESTION D'UN ARBITRE ANONYME

Une autre technique ponctuelle est suggérée par un arbitre anonyme. L'arbitre fait observer qu'une grande majorité des paires examinées dans de telles situations représentaient des non-concordances. Dans le cas des données de St. Louis, 91,5 % des paires examinées se révélèrent être des non-concordances. Compte tenu d'une telle proportion, les tendances guidant les variables de comparaison sur l'ensemble des paires reflètent principalement la situation des non-concordances. Ce raisonnement, poussé plus loin, peut amener à conclure que l'estimation des paramètres de la structure de dépendance qui sous-tend les non-concordances peut être réalisée avec succès si l'on traite l'ensemble de toutes les paires comme si c'était l'ensemble des non-concordances. L'estimation des paramètres devient triviale. Les paramètres à estimer caractérisent un modèle log-linéaire simple, sans aucune variable latente (Fienberg, Bishop et Holland, p. 24). Les paramètres descriptifs des concordances peuvent être estimés séparément par une technique itérative simple comme l'algorithme EM, combinée à de l'information disponible *a priori*.

L'approche de l'arbitre reflète, il est vrai, un modèle réaliste du processus et, en ce sens, elle est compatible avec la ligne de pensée de la présente communication. L'objectif de la communication, toutefois, est aussi d'élaborer des règles discriminantes, tout en s'en tenant à la contrainte de la structure latente. Si, dans certaines situations, la proportion de concordances est élevée ou si des liens de dépendance sont manifestes parmi les concordances, l'approche de l'arbitre n'est plus valable. Une estimation des paramètres tirée directement du modèle naturel, si elle est possible, est recommandée.

8. CONCLUSIONS

L'objectif de la recherche était de montrer comment une meilleure formulation des modèles probabilistes supportant les processus de couplage d'enregistrements peuvent contourner un pouvoir discriminant accru à la règle de couplage d'enregistrements de Fellegi-Sunter. Dans le cas des exemples de St. Louis et de Columbia, cet objectif a certainement été atteint. Le modèle enrichi exprimé en (4) est en fait le plus représentatif du processus probabiliste sous-jacent et il confère à la règle de Fellegi-Sunter un pouvoir discriminant considérablement plus élevé que le modèle à indépendance conditionnelle (2).

Les techniques appliquées aux données de St. Louis et de Columbia peuvent aussi servir à l'analyse d'autres ensembles de données créés par des processus de couplage d'enregistrements reposant sur un processus probabiliste doté d'une structure de dépendance semblable. Cette structure de dépendance apparaîtra à coup sûr dans toute

difficile. Un mot, pour terminer, sur les données de St. Louis et de Columbia. Ces données sont de très haute qualité, ce qui explique, en partie, le taux très fructueux de concordances observé tant dans l'exemple de St. Louis que dans celui de Columbia. Il serait raisonnable de s'attendre à une différence moins nette entre les diverses techniques de couplage si les données étaient de qualité inférieure.

REMERCIEMENTS

Cette communication expose les résultats généraux de recherches effectuées par le personnel du Census Bureau. Les opinions exprimées sont celles de l'auteur et ne reflètent pas nécessairement la position du Census Bureau. L'auteur est reconnaissant envers l'arbitre anonyme pour sa patience et ses suggestions constructives. L'auteur tient également à remercier William E. Winkler pour l'orientation apportée tout au long du processus d'apprentissage ayant mené à cette communication.

Les restrictions énoncées en (3) s'appliquent également ici. En outre, le modèle doit se plier à d'autres contraintes. Les contraintes suivantes sont imposées aux termes d'interaction du troisième ordre sont les suivantes:

(5)
$$\eta_{j,l}^{1,l} = -\eta_{j,l}^{1,0}; \quad \eta_{j,l}^{1,l} = -\eta_{0,l}^{1,l}.$$

L'intervalle de variation des indices est $1 \leq j < l \leq 4$. Les contraintes imposées aux termes d'interaction du troisième ordre sont les suivantes:

(6)
$$\Phi_{ij,l,m}^{j,l,l,1} = -\Phi_{ij,l,m}^{j,l,l,0}; \quad \Phi_{ij,l,m}^{j,l,l,1} = -\Phi_{ij,0,l,m}^{j,l,l,m};$$

L'intervalle de variation des indices, dans ce cas, est $1 \leq j < l < m \leq 4$. Enfin, les contraintes visant les termes d'interaction du quatrième ordre sont:

(7)
$$\begin{aligned} \Psi_{1,2,3,4}^{1,1,2,3,4} &= -\Psi_{1,2,3,4}^{1,2,3,4}; \quad \Psi_{1,2,3,4}^{1,1,2,3,4} = -\Psi_{1,2,3,4}^{1,2,3,4}; \\ \Psi_{1,2,3,4}^{1,1,2,3,4} &= -\Psi_{1,2,3,4}^{1,2,3,4}; \quad \Psi_{1,2,3,4}^{1,1,2,3,4} = -\Psi_{1,2,3,4}^{1,2,3,4}. \end{aligned}$$

Il est naturel de s'attendre à ce que le modèle enrichi (4) ait un pouvoir discriminant supérieur, puisqu'il tient compte des interactions entre les variables liées au ménage. À la section 6, les performances des deux modèles sont présentées.

4.5 Estimation des paramètres pour des modèles avec liens de dépendance

L'estimation des paramètres, dans le cas des modèles avec liens de dépendance, est beaucoup plus difficile que pour les modèles à indépendance conditionnelle. En ce qui concerne l'exemple de St. Louis, l'algorithme de notation donné par Haberman (1979, p. 547) a été employé pour l'estimation des paramètres du modèle enrichi (4). Cette technique peut être considérée comme un algorithme EM dans lequel la partie maximisation (étape M) est une application de l'algorithme de Newton-Raphson. La difficulté la plus importante que pose cette technique est le choix d'un point de départ. D'abord, les paramètres du modèle à indépendance conditionnelle (2) sont estimés au moyen de l'algorithme EM présenté à la sous-section 3.3. Puis un modèle intermédiaire est construit. Dans le présent cas, le modèle intermédiaire incorpore tous les termes d'interaction, jusqu'au deuxième ordre, du modèle enrichi (4). Les paramètres du modèle à indépendance conditionnelle déjà estimés peuvent servir à établir le point de départ permettant l'estimation des paramètres du modèle intermédiaire grâce à l'algorithme de notation. Enfin, les estimations des paramètres du modèle intermédiaire sont utilisées comme point de départ pour estimer les paramètres du modèle enrichi (4), au moyen de l'algorithme de notation.

Dans la dernière section, un modèle complexe représente un processus probabiliste sous-jacent à été formulé pour les données de St. Louis. Dans cette situation, la formulation du modèle est facile, car on dispose d'une information de suivi. En pratique, évidemment, une telle information de suivi n'est pas disponible. Il est souvent trop difficile ou trop coûteux de réaliser l'ensemble du processus de formulation du modèle et d'estimation des paramètres pour déterminer la structure du processus sous-jacent et les valeurs des paramètres. Dans l'exemple de St. Louis, l'approche ponctuelle consiste à rajuster les paramètres du processus tirés du modèle à indépendance conditionnelle (2) pour obtenir un modèle plus discriminant. Notons qu'en vertu tant du modèle (2) que du modèle (4), pour une paire formant une concordance, les correspondances ou divergences des zones de comparaison sont indépendantes. Ainsi, la formule suivante s'applique aux deux situations:

$$m(\gamma) = \prod_{i=1}^N m_i^{x_i} (1 - m_i)^{1-x_i},$$

m_i est la probabilité de correspondance de la zone i pour deux enregistrements formant une concordance. En outre, $x_i = 0$ si le profil γ prévoit une divergence à la zone i et $x_i = 1$ s'il prévoit une correspondance. L'idée de la méthode ponctuelle est de conserver la structure d'indépendance conditionnelle énoncée en (2), tout en rajustant les valeurs des m_i s. Les probabilités de correspondance, pourvu que la paire forme une concordance, évaluées selon le modèle à indépendance conditionnelle et le modèle enrichi sont données au tableau 3. L'écart entre la probabilité associée au modèle enrichi et celle associée au modèle à indépendance conditionnelle est très élevé pour certaines zones. Ainsi, cet écart est prononcé dans le cas de la zone "prénom".

Tableau 3

Probabilités de correspondance pour les paires formant des concordances

Modèle	Zone de comparaison	Indépendance conditionnelle	enrichi
Nom de famille	Prénom	.9430	.9561
	Initiale	.2125	.5222
	Numéro de maison	.9692	.9724
	Nom de rue	.9179	.9194
	Numéro de téléphone	.6619	.6887
Âge		.3903	.8602
Lien avec le répondant		.3353	.4986
Etat matrimonial		.6072	.8547
Sexe		.6134	.4842
Race		.9672	.9018

proviennent du même ménage. Par conséquent, compte tenu de la concordance des noms de famille, les probabilités de concordance des autres zones liées au ménage sont plus grandes que les probabilités marginales. La nature des liens de dépendance entre les variables liées au ménage est examinée ci-après.

4.3 Mesure des liens de dépendance

Pour construire un modèle représentatif du processus de couplage dans l'exemple de St. Louis, les liens de dépendance entre les variables liées au ménage doivent être évalués. L'information sur la variable latente le permet. Le tableau 1 donne les corrélations des réponses des comparaisons entre enregistrés pour les zones de comparaison, dans le cas des concordances. Le tableau 2 donne les corrélations des réponses des comparaisons pour les zones de comparaison, dans le cas des concordances, dans le cas des non-concordances. Dans les deux tableaux, toutes les corrélations n'est pas indiquée seulement si elle est inférieure à .01.

Les corrélations du tableau 1 sont relativement peu élevées et ne permettent pas de croire, dans l'ensemble, qu'il existe une importante structure de dépendance entre les variables de comparaison, lorsqu'on examine uniquement les concordances. Notons en particulier que les corrélations entre les variables liées au ménage sont faibles dans le cas des concordances, ce qui laisse croire qu'il existe peu ou pas de dépendance. Cela peut s'expliquer par le fait que parmi les paires formant des concordances, le taux de correspondance pour n'importe quelle zone liée au ménage est très élevé et affiche un comportement voisin de celui d'une constante.

Tableau 1
Corrélations entre certaines zones de comparaison pour l'ensemble des paires formant des liens

État	Nom de rue	N° de maison	N° de téléphone	État matrimonial
Prénom	.123	0.	.045	.032
Initiale	1	.010	.161	.079
N° de maison	.017	.194	.037	0.
Nom de rue	.01	1	.035	0.
N° de téléphone	.161	.035	1	.107
Âge	.051	.004	.075	.118
État matrimonial	.079	0.	.107	1

Au tableau 2, toutefois, les effets du groupage sont évidents, comme le montrent les valeurs élevées des corrélations associées aux variables liées au ménage quand on considère adéquatement les non-concordances. Un modèle représentant le processus probabiliste sous-jacent devrait tenir compte de ces corrélations élevées en incorporant des éléments de dépendance.

Tableau 2
Corrélations entre certaines zones de comparaison pour l'ensemble des non-liens

Nom de famille	N° de maison	Nom de rue	N° de téléphone	État matrimonial	Race
.748	.326	.642	.099	.101	
1	.400	.699	.111	.105	
Nom de rue	.400	1	.292	.086	
Âge	.104	.054	.086	.165	.024
Lien avec le répondant	.121	.068	.084	.394	.049

4.4 Modèle adapté aux données de St. Louis

Afin de pouvoir faire des inférences valides sur l'état des paires, un modèle descriptif du processus probabiliste sous-jacent doit être formulé. Le modèle à indépendance conditionnelle présenté en (2) est attrayant du fait de sa simplicité. Toutefois, il est clair à ce stade que ce modèle ne représente pas correctement le processus probabiliste qui sous-tend le processus de couplage de St. Louis. Nous présentons maintenant un modèle enrichi, qui met à contribution l'information obtenue sur les liens de dépendance entre les variables liées au ménage.

Pour que la structure plus générale du modèle enrichi puisse être comprise, certaines conventions doivent être établies quant à l'indexation des zones de comparaison: la zone de comparaison 1 est le nom de famille, la zone de comparaison 2 est le numéro de maison, la zone de comparaison 3 est le nom de rue et la zone de comparaison 4 est le numéro de téléphone. Les sept autres zones de comparaison sont indexées arbitrairement par les valeurs 5 à 11. Le modèle enrichi tient compte de tous les effets d'interaction possibles entre les zones 1 à 4 parmi les non-concordances. La représentation log-linéaire du modèle enrichi est la suivante:

$$\log(v_{k,i_1,\dots,i_{11}}) = \mu + \lambda_k + \sum_{i_1=1}^4 \alpha_{i_1}^k + \sum_{i_2=1}^4 \beta_{i_2}^{k,i_1} + \sum_{\substack{\{j \leq l \leq 4\} \\ \{i_j, i_l\}}} \gamma_{i_j, i_l}^{k,i_1} + \sum_{\substack{\{1 \leq j < l < m \leq 4\} \\ \{i_j, i_l, i_m\}}} \Phi_{i_j, i_l, i_m}^{k,i_1} + \Psi_{i_1, i_2, i_3, i_4}^{k,i_1} \quad (4)$$

Notons le coefficient $(1 - k)$ qui multiplie les termes d'interaction des variables liées au ménage, indiquant que le lien de dépendance entre les variables liées au ménage ne concerne que les non-concordances. On peut remarquer le contraste avec la symétrie du modèle à indépendance conditionnelle présenté en (2).

Pour cette application particulière, la variable latente est rendue observable grâce à une étude de suivi approfondie, effectuée pour les besoins de la présente recherche et d'autres travaux. Dans le cas qui nous occupe, l'information extraite de la variable latente mène à la construction d'un modèle représentatif du processus probabiliste sur lequel repose le processus de couplage des enregistrements. Enfin, le pouvoir discriminant de ce modèle est comparé à celui du modèle à indépendance conditionnelle. Les motivations à la base de la construction du modèle sont présentées dans les sous-sections qui suivent.

4.2 Groupage et liens de dépendance

Le but du couplage d'enregistrements est de trouver autant de concordances que possible, compte tenu d'une limite maximum imposée à l'erreur de type I. Le premier obstacle est souvent la taille des fichiers. Ceux-ci peuvent être très volumineux, ce qui rend impossible l'examen de toutes les paires formées d'un enregistrement du fichier A et d'un enregistrement du fichier B. Le groupage est une solution qui est envisagée lorsqu'un examen exhaustif des paires est une activité qui serait trop coûteuse ou trop longue. Le principe du groupage est le suivant : pour réduire le nombre de comparaisons et d'autres opérations connexes, les enregistrements de chaque fichier sont attribués à des groupes en fonction de la valeur de quelques caractéristiques clés. Ces caractéristiques sont appelées les variables de groupage. Seuls les enregistrements dont les variables de groupage ont les mêmes valeurs peuvent être réunis en paires. Puisque les enregistrements constituant une concordance sont en général identiques au niveau des caractéristiques de groupage, il est naturel de s'attendre à ce que la vaste majorité des paires omises ne représentent pas des concordances, compte tenu du schéma de groupage.

Dans l'exemple de St. Louis, le fichier du recensement compte 15,048 enregistrements, tandis que le fichier PBS contient 12,072 enregistrements. Il y a donc une possibilité de plus de 180,000,000 paires à examiner. Ce nombre est trop élevé et le groupage doit être utilisé pour garder le problème à une dimension raisonnable. Par conséquent, les enregistrements sont groupés selon le premier caractère du nom de famille et une unité géographique appelée géocode. La zone géographique couverte par un géocode particulier peut comprendre plusieurs pâtés de maisons, ou encore deux rues voisines (ou plus), perpendiculaires ou parallèles. Cette répartition donne des groupes de taille raisonnable. Selon ce schéma, 116,305 paires fournissent l'information permettant d'établir les inférences.

Malheureusement, bien qu'il réduise l'ampleur du problème, le groupage selon les géocodes s'accompagne d'un effet secondaire indésirable : il entraîne la présence de forts liens de dépendance entre les variables liées au ménage, c'est-à-dire le nom de famille, le numéro de maison, le nom de rue et le numéro de téléphone. Prenons par exemple deux individus formant une non-concordance, mais qui font partie du même groupe. Supposons en outre que ces deux individus aient le même nom de famille. Intuitivement, compte tenu de cette information, on peut croire que les chances sont plus grandes que les deux individus

Le processus d'estimation des paramètres doit être assez fiable pour empêcher que l'erreur d'estimation n'entraîne une perte du pouvoir discriminant.

Une caractéristique des modèles à classes latentes fait qu'ils sont enclins à l'erreur d'estimation : ils sont non identifiabiles, en ce sens que les équations maximisant la vraisemblance admettent plus d'une solution. L'estimation des paramètres pour des modèles non identifiabiles demeure difficile et déroutante. Toutefois, l'auteur a constaté par expérience que dans le cas des modèles à indépendance conditionnelle, le fait d'être non identifiabiles n'est pas habituellement un facteur déterminant de l'erreur d'estimation. Une part plus grande de l'erreur est généralement attribuable à l'incapacité du modèle à représenter authentiquement le processus probabiliste sous-jacent.

Une technique appropriée d'estimation des paramètres dans le cas des modèles à indépendance conditionnelle consiste à considérer le problème comme étant celui de trouver un estimateur du maximum de vraisemblance dans une situation d'"observations manquantes". Dans le cas présent, l'observation manquante est la variable latente, c.-à-d. l'état de chaque paire. Dans le contexte général de l'estimation de paramètres en situation d'"observations manquantes", les algorithmes "espérance-maximisation" (EM) sont très populaires. En fait, l'algorithme EM s'applique sans difficulté à l'estimation des paramètres du modèle à indépendance conditionnelle donné en (2) (Winkler 1988). Toutefois, s'il y a un écart considérable par rapport à l'hypothèse d'indépendance, la valeur des estimations devient difficile à interpréter. (Un exemple d'une telle situation est donné à la section 4.)

4. LES DONNÉES DE ST. LOUIS : UN EXEMPLE DE PROCESSUS DE COUPLAGE COMPLEXE

La présente section expose un exemple particulier de processus de couplage d'enregistrements. Un modèle est élaboré expressément pour représenter le processus probabiliste qui sous-tend ce processus de couplage. L'on s'attend à ce que ce modèle confère un pouvoir discriminant plus élevé à l'application de la règle de Fellegi-Sunter que ne le ferait le modèle à indépendance conditionnelle.

4.1 Variable latente observable

Notre exemple se fonde sur des données recueillies en 1988 au cours d'une répétition générale précédant la réalisation du recensement décennal. Essentiellement, il y a deux relevés distincts et supposément exhaustifs de la population habitant une zone géographique définie de la ville de St. Louis, au Missouri. Pour chaque relevé et pour chaque individu observé au moment du relevé, un enregistrement est créé et diverses caractéristiques de l'individu sont inscrites. Ces caractéristiques sont les suivantes : numéro de maison, numéro de téléphone, nom de rue, prénom, nom de famille, initiale, état matrimonial, âge, race, sexe, lien avec le répondant. Les enregistrements des deux enquêtes sont couplés les uns avec les autres.

3. MODÈLES POUR LE COUPLAGE D'ENREGISTREMENTS

Deux modèles représentant les processus probablistes sous-jacents sont présentés dans cette section. Le premier modèle est une formulation générale de n'importe quel processus sous-jacent. Le deuxième modèle est une application du premier. Dans certaines situations, le deuxième modèle est une bonne représentation du processus probabliste sous-jacent et la règle de Fellegi-Sunter, selon ce modèle, est la plus discriminante. L'évaluation des paramètres est examinée, de façon que les expressions intervenant dans la règle de Fellegi-Sunter puissent être évaluées.

3.1 Modèles à classes latentes

En raison de la nature particulière du processus de couplage d'enregistrements, le processus probabliste sous-jacent peut toujours être représenté par un modèle à classes latentes. Un tel modèle s'articule autour de variables latentes. En termes généraux, une variable latente est une variable non observable, qui caractérise toute observation associée au processus probabliste. Les variables latentes permettent de classer les observations dans des classes latentes. Dans le cas qui nous occupe, les observations sont les vecteurs de comparaison (c.-à-d. les profils de comparaison). Une variable latente évidente, permettant de classer les observations dans deux classes latentes, est l'état de la paire associée à chaque vecteur de comparaison. Cet état peut être une concordance ou une non-concordance. Les classes latentes correspondantes sont la classe des concordances et la classe des non-concordances. Une représentation mathématique est présentée ci-dessous, pour permettre l'élaboration de modèles à classes latentes particuliers.

Soit w_{k,i_1, \dots, i_N} le nombre de paires ayant les attributs suivants: si $k = 0$ les paires correspondantes sont classées comme non-concordantes, et si $k = 1$ elles sont classées comme concordantes. En outre, si $i_s = 0$, les paires correspondantes n'affichent pas de concordance des enregistrements au niveau de la zone de comparaison s , et si $i_s = 1$, les paires affichent une concordance des enregistrements au niveau de la zone de comparaison s . Notons que $s = 1, \dots, N$, où N est le nombre de zones de comparaison. Il importe de se rappeler que les nombres w_{k,i_1, \dots, i_N} ne peuvent être observés. Ce qui peut être observé, plutôt, ce sont les nombres globaux pour l'ensemble des classes latentes. Les nombres globaux sont dénotés par v_{i_1, \dots, i_N} où

$$(1) \quad v_{i_1, \dots, i_N} = v_{0,i_1, \dots, i_N} + v_{1,i_1, \dots, i_N}.$$

Bien que seuls les nombres globaux soient observables dans des opérations de couplage d'enregistrements, les modèles sont habituellement exprimés en termes de nombres de base. Une telle formulation est faite uniquement à des fins de commodité. La sous-section qui suit présente un cas plus précis de modèle à classes latentes simple pour le couplage d'enregistrements.

3.2 Indépendance conditionnelle

Les modèles à indépendance conditionnelle sont les modèles à classes latentes les plus simples. Malgré leur simplicité, ces modèles offrent une représentation exacte du processus probabliste sous-jacent dans certaines situations. Goodman (1974) effectue une analyse approfondie de plusieurs modèles à indépendance conditionnelle. Haberman (1979) présente des techniques appropriées d'estimation des paramètres. Dans cette section, le modèle à indépendance conditionnelle pour le couplage d'enregistrements est présenté, et ses conséquences du point de vue du processus probabliste sous-jacent sont énoncées. Le modèle est le mieux décrit par sa représentation log-linéaire:

3.3 Estimation des paramètres pour le modèle à indépendance conditionnelle

Une fois qu'un modèle a été formulé, les valeurs de ses paramètres doivent être déterminées. Ensuite, la règle de Fellegi-Sunter est établie d'après le modèle, une fois connues les valeurs estimatives correspondantes de $m(\gamma)$ et $n(\gamma)$.

l'élément déterminant du processus de couplage d'enregistrements. $m(\gamma)$ et $n(\gamma)$ jouent un rôle crucial dans la construction des règles de couplage et en particulier des règles ayant le pouvoir discriminant le plus élevé. Les règles de couplage sont un moyen d'obtenir des concordances. Elles sont définies dans la section qui suit.

2.3 Règles de couplage d'enregistrements

En pratique, une règle de couplage d'enregistrements permet de classer les paires produites par un processus de couplage d'enregistrements en trois catégories possibles :

un lien, un non-lien et un lien possible. Un lien est établi quand une concordance est déduite, et un non-lien est établi quand une non-concordance est déduite. Les paires classées comme des liens possibles sont mises de côté en vue d'un examen plus approfondi et seront finalement classées comme des liens ou des non-liens. La règle est fondée uniquement sur la valeur des vecteurs de comparaison correspondant à chaque paire. Les erreurs associées à une règle de couplage d'enregistrements sont de deux types : l'erreur de type I mesure la proportion des non-concordances parmi les paires classées comme des liens en vertu de la règle de couplage, et l'erreur de type II mesure la proportion des concordances parmi les paires classées comme des liens.

L'objectif du couplage d'enregistrements, du point de vue traité dans cette communication, est de construire la règle de couplage ayant le pouvoir discriminant le plus élevé, c'est-à-dire celle qui permettra d'établir un maximum de liens tout en gardant le plus bas possible l'erreur de type I. À cette fin, indexons les profils de comparaison selon les valeurs décroissantes de $m(\gamma)/n(\gamma)$ pour obtenir la séquence $\{\gamma_1, \gamma_2, \dots, \gamma_m\}$, où M est le nombre total de paires. Fellegi et Sunter (1969) montrent que la règle qui déclare comme "liens" les paires dont l'indice est inférieur à une limite supérieure K est la règle de couplage d'enregistrements la plus discriminante. La limite supérieure K est fonction de l'erreur de type I maximum tolérée. La règle est la plus discriminante en ce sens que pour la même tolérance en ce qui a trait à l'erreur de type I, il est impossible de trouver une autre règle qui, après un nombre élevé de comparaisons, trouvera plus de concordances. Ce résultat est une application directe du lemme de Neyman-Pearson (DeGroot 1986, p. 444-445). Deux utilisations de la règle de Fellegi-Sunter sont illustrées à la section 6.

La règle de couplage de Fellegi-Sunter se fonde sur le rapport $m(\gamma)/n(\gamma)$. Habituellement, ce ratio est estimé à partir des données, par l'entremise d'un modèle du processus probabiliste sous-jacent. On suppose que le modèle est une représentation authentique du processus probabiliste. S'il ne s'agit pas d'une représentation authentique, l'emploi de $m(\gamma)/n(\gamma)$ dans la règle de Fellegi-Sunter ne donnera pas nécessairement la règle de couplage la plus discriminante. Il faut donc veiller à choisir avec soin le modèle. La section qui suit présente des modèles aptes à décrire le processus probabiliste sous-jacent dans certaines situations données.

À la section 5, nous présentons une autre méthode possible de construction de modèles probabilistes approximatifs. Le modèle produit par cette méthode est comparé à ceux présentés aux sections 3 et 4, sous l'angle du pouvoir discriminant des règles de couplage découlant des modèles. Les résultats des comparaisons sont présentés à la section 6. À la section 7, les suggestions d'un arbitre anonyme visant à améliorer la méthode exposée dans cette communication sont présentées. À la section 8, des conclusions sont tirées et des lignes directrices sont énoncées.

2. LE MODÈLE DE FELLEGI-SUNTER D'ENREGISTREMENTS POUR LE COUPLAGE

2.1 Processus de couplage d'enregistrements

La présente communication s'intéresse à la construction de nouvelles techniques de couplage d'enregistrements. Toutefois, avant d'examiner plus à fond ces nouvelles méthodes, certaines notions de base doivent être rappelées. Le concept de "processus de couplage d'enregistrements" doit d'abord être passé en revue. Considérons deux fichiers, le fichier A et le fichier B, contenant tous deux des enregistrements, lesquels enregistrements représentent chacun un individu. Un processus de couplage d'enregistrements réunit un enregistrement du fichier A et un enregistrement du fichier B. Les enregistrements sont comparés, ce qui produit le profil de comparaison γ . Pour les besoins de cette communication, le profil de comparaison est un vecteur $\gamma = [\gamma_1, \dots, \gamma_N]$, où N est le nombre de dimensions du vecteur. Chaque dimension correspond à une zone de comparaison observée pour chaque individu (nom de famille, âge, adresse, etc.). Sans perte de généralité, γ_i reçoit la valeur 0 si les contenus de la zone de comparaison i ne concordent pas, et la valeur 1 s'ils concordent. L'espace de comparaison T est défini comme l'ensemble de tous les vecteurs binaires (c.-à-d. formés d'éléments valant 0 ou 1) de dimension N .

2.2 Processus probabilistes sous-jacents

Un processus de couplage d'enregistrements est régi par un processus probabiliste sous-jacent. Une bonne connaissance du processus probabiliste est nécessaire pour qu'on puisse extraire de l'information du processus de couplage d'enregistrements. La formulation du processus probabiliste sous-jacent est présentée ici en termes généraux. Un examen plus détaillé est effectué dans la section suivante. Considérons un profil de comparaison particulier γ , et définissons $m(\gamma)$ comme la probabilité d'observer γ , dans l'hypothèse où les deux enregistrements produisant γ , s'ils sont réunis, représentent le même individu. De même, définissons $n(\gamma)$ comme la probabilité d'observer γ , dans l'hypothèse où les deux enregistrements produisant γ , s'ils sont réunis, ne représentent pas le même individu. Ces deux probabilités conditionnelles, de même que la probabilité d'une concordance, définissent le processus probabiliste sous-jacent. Le processus probabiliste est

Le pouvoir discriminant des structures de dépendance dans le couplage d'enregistrements

YVES THIBAUDEAU¹

RÉSUMÉ

Dans un processus de couplage d'enregistrements, des enregistrements provenant de deux fichiers sont réunis en paires, formées d'un enregistrement de chacun des fichiers, à des fins de comparaison. Chaque enregistrement représente un individu. Une paire ainsi formée est une "concordance" si les deux enregistrements représentent le même individu. Une paire est une "non-concordance" si les deux enregistrements ne représentent pas le même individu. Le processus de couplage d'enregistrements repose sur un processus probabiliste. Une règle de couplage déduit l'état (concordance ou non) de chaque paire d'enregistrements d'après la valeur de la comparaison. La paire est déclarée un "lien" si une concordance est déduite, et un "non-lien" si une non-concordance est déduite. Le pouvoir discriminant d'une règle de couplage est la capacité de la règle de désigner un nombre maximum de concordances comme des liens, tout en gardant au minimum le taux de non-concordances désignées comme des liens. En général, pour construire une règle de couplage discriminante, il faut faire certaines hypothèses quant à la structure du processus probabiliste sous-jacent. Dans la majorité de la documentation existante, il est supposé que le processus probabiliste sous-jacent est une manifestation du modèle à classes latentes avec indépendance conditionnelle. Toutefois, dans bien des situations, cette hypothèse est fautive. En fait, de nombreux processus probabilistes sous-jacents n'affichent pas les caractéristiques clés associées aux modèles à classes latentes avec indépendance conditionnelle. Cette communication présente des modèles plus généraux. En particulier des modèles à classes latentes avec liens de dépendance sont étudiés, et nous montrons comment ils peuvent améliorer le pouvoir discriminant de règles de couplage particulières.

MOTS CLÉS : Règle de couplage d'enregistrements; modèle à classes latentes; algorithmes "espérance-maximisation".

1. INTRODUCTION

Le but de la présente communication est de montrer comment le pouvoir discriminant de règles de couplage d'enregistrements peut être accru lorsque des modèles probabilistes plus représentatifs des processus probabilistes sous-jacents sont formulés. À cette fin, un exemple particulier de couplage d'enregistrements est présenté, et le modèle à indépendance conditionnelle, traditionnellement employé pour le couplage d'enregistrements, est comparé à un modèle plus descriptif, c'est-à-dire un modèle permettant l'expression de liens de dépendance plus complexes entre certaines des variables concernées.

En premier lieu, une certaine terminologie doit être passée en revue. À la section 2, la définition d'un processus de couplage d'enregistrements est énoncée, et une formulation générale du processus probabiliste qui sous-tend le processus de couplage est présentée. Cette formulation mène à l'expression de deux concepts centraux: le concept de règle de couplage d'enregistrements et celui de règle de couplage la plus discriminante.

À la section 3, des modèles probabilistes pour le couplage d'enregistrements sont examinés. Dans la première partie de la section 3, la famille des modèles à classes latentes est présentée, et nous montrons comment

cette famille offre des modèles naturels pour le processus probabiliste sous-jacent à un processus de couplage d'enregistrements. Dans la deuxième partie, l'attention est centrée sur un modèle particulier de la famille des modèles à classes latentes: le modèle à classes latentes avec indépendance conditionnelle. Ce modèle est intéressant du fait qu'il se prête facilement à des calculs informatiques. Dans la troisième partie, des techniques d'inférence adaptées au modèle à indépendance conditionnelle sont décrites.

À la section 4, une application est présentée. Dans cette application, le vrai et le faux sont connus, c'est-à-dire que nous savons quelles sont les véritables concordances et non-concordances. Dans la première partie, la façon dont l'information sur le vrai et le faux a été obtenue est expliquée. La deuxième partie montre comment des liens de dépendance entre les zones de comparaison sont produits. Dans la troisième partie de la section 4, la connaissance du vrai et du faux est utilisée pour évaluer les liens de dépendance entre les zones de comparaison. On en arrive alors à la quatrième partie, qui est la formulation d'un modèle plus représentatif de la structure probabiliste sur laquelle repose le processus de couplage d'enregistrements. La dernière partie comporte une brève description des techniques d'estimation des paramètres pour des modèles à classes latentes généralisés.

¹ Yves Thibaudreau, U.S. Bureau of the Census, Federal Bldg. 4, Room 3000, Washington, D.C. 20233.

- WILLIAMS, B.C., DEMITRACK, L.B., et FRIES, B.E. (1992). The accuracy of the National Death Index when personal identifiers other than Social Security Number are used. *American Journal of Public Health*, 82, 1145-1147.
- WINKLER, W.E. (1985a). Preprocessing of lists and string comparison. Dans *Record Linkage Techniques - 1985*, (Eds. W. Alvey et B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 181-187.
- WINKLER, W.E. (1985b). Exact matching lists of businesses: blocking, subfield identification, and Information Theory. Dans *Record Linkage Techniques - 1985*, (Eds. W. Alvey et B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 227-241.
- WINKLER, W.E. (1991). Documentation of record-linkage software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- WINKLER, W.E., et THIBAUDEAU, Y. (1992). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

- HOWE, G.R., et LINDSAY, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers in Biomedical Research*, 14, 327-340.
- HOWE, G.R., et SPASOFF, R.A. (Eds.) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press.
- JABINE, T.B., et SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.
- JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.
- JOHANSEN, H.T. (1986). Record linkage of national surveys: The Nutrition Canada example. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 153-163.
- JOHNSON, E.G., et TURKEY, J.W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson et Tsao Data. Dans *Design, Data and Analysis*, (Ed. C.L. Mallows) New York: John Wiley and Sons.
- JOHNSON, R.A. (1991). Methodology for Evaluating Errors in U.S. Department of Justice Attorney Workload Data. Report techniques non-publié, General Accounting Office, Washington, D.C.
- KETLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.
- KERSHAW, D., et FAIR, J. (1979). *The New Jersey Income and Maintenance Experiment: Operations, Surveys, and Administration*, Volume I. New York: Academic Press.
- KILSS, B., et ALVEY, W. (Eds.) (1984a). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. I, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., et ALVEY, W. (Eds.) (1984b). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. II, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., et ALVEY, W. (Eds.) (1984c). *Statistics of Income and Related Administrative Record Research: 1984*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., et ALVEY, W. (Eds.) (1987). *Statistics of Income and Related Administrative Record Research: 1986-1987*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., et JAMERSON, B. (Eds.) (1990). *Statistics of Income and Related Administrative Record Research 1988-1989*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.
- KILSS, B., et SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin*, Vol. 41, 10, 14-22.
- LAPLANT, W. (1988). User's Guide for the Generalized Record Linkage Program Generator (GENLINK). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- LAPLANT, W. (1989). User's Guide for the Generalized Address Standardizer (GENSTAN). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.
- NEWCOMBE, H.B., FAIR, M.E., et LALONDE, P. (1992). The use of names for linking personal records (avec discussion). *Journal of the American Statistical Association*, 87, 1193-1208.
- NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., et ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado Uranium Workers. *Computers in Biology and Medicine*, 13, 157-169.
- NICHOLL, J.P. (1986). The use of hospital in-patient data in the analysis of the injuries sustained by road accident casualties. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 243-244.
- PALETZ, D. (1989). Name standardization software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- ROGOT, E., SORLIE, P.D., et JOHNSON, N.J. (1986). Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Disease*, 39, 719-734.
- ROGOT, E., SORLIE, P.D., JOHNSON, N.J., GLOVER, C.S., et TREASURE, D.W. (1988). A Mortality Study of One Million Persons. Public Health Service, National Institutes of Health, Washington, D.C.
- SCHIRM, A.L., et PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (avec discussion). *Journal of the American Statistical Association*, 82, 965-990.
- SMITH, M.E., et NEWCOMBE, H.B. (1975). Methods for computer linkage of hospital admission-separation records for cumulative health histories. *Methods of Information in Medicine*, 14, 118-125.
- TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Section on Statistical Computing, American Statistical Association* 283-288.
- WENTWORTH, D.N., NEATON, J.D., et RASMUSSEN, W.L. (1983). An evaluation of the Social Security Administration Master Beneficiary Record and the National Death Index in the ascertainment of vital status. *American Journal of Public Health*, 73, 1270-1274.

- BELIN, T.R. (1991). Using mixture models to calibrate error rates in record-linkage procedures, with application to computer-matching for census undercount estimation. These de doctoral, Department of Statistics, Harvard University. (Publié par University Microfilms, Inc.)
- BELIN, T.R., et RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, Washington, D.C.
- BOYLE, C.A., et DECOUFLÉ, P. (1990). National sources of vital status information: Extent of coverage and possible selectivity in reporting. *American Journal of Epidemiology*, 131, 160-168.
- BROWN, P., LAPLANT, W., LYNCH, M., ODELL, S., THIBAUDEAU, Y., et WINKLER, W. (1988). Collective Documentation for the 1988 PES Computer Match Processing and Printing. Vols. I-III, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.
- BUREAU OF THE CENSUS (1988-1991). 1990 Decennial Census Information Memorandum Series, Decennial Planning Division, Bureau of the Census, Washington, D.C.
- [Nota: Tous les rapports de la série mentionnée ci-dessus sont accompagnés de l'avertissement suivant: "Ces documents ont été rédigés à l'intention des divisions de la planification et de l'exploitation du Census Bureau; ces divisions connaissent bien les données de base, les expériences antérieures, la terminologie et les procédés de même que le cadre général du recensement décennal, ses objectifs et la relation mutuelle des opérations et des systèmes. Ces documents ne sont donc PAS [soullignement tiré de l'original] destinés à la diffusion et ne doivent pas sortir du Census Bureau sans l'autorisation préalable de Jim Dinwiddie ([301]-763-5270) de la division de la planification du recensement décennal."]
- BUREAU OF THE CENSUS (1987-1991). STSD Decennial Census Memorandum Series, Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.
- CARPENTER, M., et FAIR, M.E. (Éds.) (1990). Canadian Epidemiology Research Conference - 1989: *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Centre for Health Information, Statistique Canada, Ottawa, Ontario.
- CHERNOFF, H. (1980). The identification of an element of a large population in the presence of noise. *Annals of Statistics*, 8, 1179-1197.
- CHILDERS, D. (1989). 1990 PES Within Block Matching - Clerical Matching Group. STSD Decennial Census Memorandum Series #V-69, U.S. Bureau of the Census, Washington, D.C.
- CITRO, C.F., et COHEN, M.L. (Éds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C.: National Academy Press.
- COHEN, M.L. (1990). Adjustment and reapportionment - Analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.
- COOMBS, J.W., et SINGH, M.P. (Éds.) (1988). *Recueil: Symposium sur les utilisations statistiques des données Administratives*, Statistique Canada, Ottawa, Ontario.
- COPAS, J., et HILTON, F. (1990). Record Linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A*, 153, 287-320.
- CURR, J.D., FORD, C.E., PRESSER, S., PALMER, M., BABCOCK, C., et HAWKINS, C.M. (1985). Ascertainment of vital status through the National Death Index and the Social Security Administration. *American Journal of Epidemiology*, 121, 754-766.
- DANIEL, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1, 311-341.
- DEMPSTER, A.P., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- DONOGHUE, G. (1990). Clerical Specifications for the 1990 Post Enumeration Survey Before Followup Matching - Special Matching Group. STSD Decennial Census Memorandum Series #V-92, U.S. Bureau of the Census, Washington, D.C.
- DULBERG, C.S., SPASOFF, R.A., et RAMAN, S. (1986). Reactor clean-up and bomb test exposure study. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Éds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 59-62.
- ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and Beyond (avec discussion). *Journal of the American Statistical Association*, 80, 98-131.
- ERICKSON, E.P., KADANE, J.B., et TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, 84, 927-944.
- FAGERLIND, I. (1975). *Formal Education and Adult Earnings: A Longitudinal Study on the Economic Benefits of Education*. Stockholm: Almqvist and Wiksell.
- FELLIG, I.P., et SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (avec discussion). *Statistical Science*, 1, 1-39.
- GOLDACRE, M.J. (1986). The Oxford record linkage study: Current position and future prospects. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press, 106-129.
- HILL, T. (1981). Generalized Iterative Record Linkage System: GIRLS. (Glossary, Concepts, Strategy Guide, User Guide). Division du développement de système, Statistique Canada, Ottawa, Ontario.
- HILL, T., et PRING-MILL, F. (1986). Generalized iterative record linkage system: GIRLS, (édition révisée). Division du développement de système, Statistique Canada, Ottawa, Ontario.
- HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, 46, 261-269.

Tableau 4.6

Taux d'erreur d'appariement moyen pour diverses combinaisons de traitements (moyenne calculée pour trois régions d'essai et cinq seuils d'exclusion: (70%, 72.5%, 75%, 77.5% et 80% du fichier de l'EP ayant pu être couplés)

Niveaux des facteurs formant la combinaison (A B C D E F G H)	Taux d'erreur d'appariement moyen pour trois régions d'essai et cinq seuils d'exclusion				
	70%	72.5%	75%	77.5%	80%
3 3 2 3 2 1 1	0.00493	0.00154	0.00137	0.00161	0.00124
3 3 2 3 1 2 1	0.00191	0.00153	0.00138	0.00156	0.00155
3 3 1 3 2 1 2	0.00153	0.00138	0.00156	0.00155	0.00155
3 3 1 3 1 2 1	0.00153	0.00138	0.00156	0.00155	0.00155
3 3 1 3 2 1 1	0.00153	0.00138	0.00156	0.00155	0.00155
3 3 1 3 1 1 2	0.00191	0.00153	0.00138	0.00156	0.00155
2 3 2 3 1 2 1	0.00124	0.00161	0.00137	0.00154	0.00154
3 3 2 3 2 2 2	0.00137	0.00154	0.00154	0.00154	0.00154
3 3 1 3 1 2 1	0.00154	0.00154	0.00154	0.00154	0.00154
3 3 2 3 2 1 1	0.00493	0.00154	0.00137	0.00161	0.00124

À des fins de comparaison, nous présentons dans le

tableau 4.6 le rendement moyen de quelques-unes des combinaisons de traitements susceptibles de figurer parmi les plus efficaces. Ainsi donc, il semble que les meilleures combinaisons après (2,3,3,1,2,1,2) soient (3,3,1,3,2,2,2,2) et (3,3,1,3,2,1,2,1). Cellules-ci se distinguent de la première par les points suivants: attribution de poids d'une valeur de ± 6 pour les zones de nom, attribution de poids d'une valeur préétablie pour les zones autres que les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance imparfaite dans les zones de nom comme dans les autres zones, utilisation de la forme normalisée du prénom, et inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement. Ces deux mêmes combinaisons diffèrent entre elles sur deux points: la première prévoit la correction des poids composites pour les cas de concordance corrélée ainsi que l'utilisation des sept chiffres du numéro de téléphone alors que la seconde ne prévoit pas de correction pour la concordance corrélée et exige l'utilisation des quatre derniers chiffres du numéro de téléphone. Les combinaisons qui prévoient l'utilisation de la pondération basée sur la fréquence pour les zones de nom ne sont pas aussi efficaces que les combinaisons qui prévoient l'attribution de poids déterminés de façon ponctuelle.

La combinaison de traitements qui a été utilisée dans les opérations d'appariement informatisé de l'EP de 1990 ressemblait beaucoup à la combinaison (5,3,2,3,2,2,2,1). En ce qui concerne les ensembles de données de précédemment analysés ici, cette combinaison a produit un taux d'erreur d'appariement moyen (pour les cinq seuils d'exclusion) de 0.00179.

4.6 Conclusions

Tandis que les résultats de cette étude tentent de mettre en balance le nombre d'enregistrements qui appartiennent à des concordances désignées et le taux d'erreur d'appariement,

Cette étude a été réalisée en majeure partie pendant que l'auteur travaillait à la division du couplage des enregistrements du Bureau of the Census des E.-U. à Washington, DC. L'auteur tient à exprimer toute sa gratitude à Don Rubin, Bill Winkler et Alan Zaslavsky ainsi qu'à l'arbitre anonyme pour les échanges fructueux auxquels ils ont participé et les commentaires utiles qu'ils ont formulés. L'auteur tient à souligner aussi l'aide qu'il a reçue grâce aux conventions sur la statistique n° 88-02 et 89-07 pendant qu'il poursuivait des études de doctorat à l'Université Harvard.

REMERCIEMENTS

un arbitre anonyme a souligné que "(...) chaque gain réalisé grâce à une meilleure méthode de couplage d'enregistrements doit être mis en parallèle avec le coût de mise en oeuvre de cette méthode." Cela est un autre arbitrage qui mérite d'être considéré par le praticien. Nous espérons que les résultats qui ont été présentés dans cette étude sur l'importance relative de divers facteurs dans le couplage d'enregistrements orienteront le travail des personnes qui élaboreront et appliqueront des logiciels de couplage. Comme certains résultats peuvent dépendre de caractéristiques particulières des données du recensement et de l'EP qui sont apparues, on peut se demander quel genre de rapport existe entre ces résultats et d'autres cadres de couplage d'enregistrements. Mais comme nous l'avons souligné au départ, l'étude du couplage d'enregistrements doit se faire idéalement à l'aide de méthodes expérimentales pour les besoins de la généralisation. Les études empiriques fondées sur des plans d'expérience constituent une source d'orientation authentique et éprouvée qui permet, par un cadre bien défini, d'élargir les connaissances des spécialistes du couplage d'enregistrements.

ABBATT, J.D. (1986). A cohort study of eldorado uranium workers. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 51-57.

ACHESON, E.D. (1967). *Medical Record Linkage*. Oxford: Oxford University Press.

ACHESON, E.D. (Ed.) (1968). *Record Linkage in Medicine*. Edinburgh: E. & S. Livingstone.

BALDWIN, J.A., ACHESON, E.D., et GRAHAM, W.J. (Eds.) (1987). *A Textbook of Medical Record Linkage*. Oxford: Oxford University Press.

BELIN, T.R. (1989a). Outline of procedure for evaluating computer matching in a factorial experiment. Note de service non-publiée, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1989b). Results from evaluation of computer matching. Note de service non publiée, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BIBLIOGRAPHIE

4) Les poids de "Fellegi-Sunter" pour les zones autres que les zones de nom donnent de meilleurs résultats que les poids déterminés de façon ponctuelle lorsqu'il n'y a pas de correction de poids pour les cas de concordance corrélée, alors que c'est l'inverse lorsqu'on effectue une telle correction. (Cependant, d'après la méthode décrite dans Belin (1991) pour estimer le niveau de bruit sous-jacent, ce phénomène ne devrait pas nécessairement être observé dans d'autres régions d'essai.)

4.5 Quelle combinaison de traitements donne les meilleurs résultats?

Pour conclure l'analyse des résultats expérimentaux, nous allons chercher à déterminer les combinaisons de traitements qui sont le plus efficaces. Pour mesurer le rendement d'une combinaison de traitements donnée, nous considérons la valeur moyenne de la variable de résultat pour les trois régions d'essai données. La variable de résultat étudiée est le taux d'erreur d'appariement pour diverses proportions des enregistrements du fichier de l'EP qui peuvent être couplés (60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5% et 90%). Les résultats pertinents figurent dans le tableau 4.5.

Tableau 4.5
Les meilleurs combinaisons de traitements pour treize seuils d'exclusion d'après l'expérience factorielle

Seuil d'exclusion	Niveaux des facteurs formant les meilleurs combinaisons (A B C D E F G H)	Taux d'erreur d'appariement moyenne pour les trois régions d'essai	
		Taux d'erreur	d'appariement
60%	3 3 2 3 2 1 1 1	0.0042	0.0047
62.5%	3 3 1 3 1 2 1 1	0.0047	0.0052
65%	3 3 1 3 2 2 2 2	0.0052	0.0071
67.5%	3 3 2 3 2 2 1 1	0.0071	0.0079
70%	2 3 2 3 1 2 1 2	0.0079	0.0081
72.5%	5 2 2 3 1 1 1 2	0.0081	0.00112
75%	5 1 1 3 2 1 2 1	0.00112	0.00133
77.5%	3 3 1 3 2 1 2 1	0.00133	0.00188
80%	2 3 2 3 1 2 1 2	0.00188	0.00571
82.5%	3 3 1 3 2 2 1 1	0.00571	0.01556
85%	5 1 2 3 2 2 1 2	0.01556	0.03023
87.5%	2 3 2 3 1 2 1 2	0.03023	0.05174
90%	2 3 2 3 1 2 1 2	0.05174	

Ces résultats contrastent avec ce que nous avons observé plus tôt; en effet, ils donnent à penser que la méthode de pondération basée sur la fréquence pour les zones de nom (niveau 5 du facteur A) est supérieure, en moyenne, à la méthode de pondération ponctuelle qui attribue des poids d'une valeur de ± 6 aux zones de nom (niveau 3 du facteur A). Cet état de fait serait imputable

apparemment à certains effets d'interaction. Lorsque la méthode ponctuelle est utilisée conjointement avec les niveaux appropriés des autres facteurs, elle semble être au moins aussi efficace que la méthode basée sur la fréquence. Nous remarquons aussi que la combinaison de traitements 2-1 pour les facteurs F et G respectivement n'est pas toujours la meilleure combinaison possible dans leur cas, malgré ce que nous avons constaté plus haut. Seul le traitement 3 du facteur D (le volet qui correspond à l'utilisation de la méthode de Winkler pour les cas de concordance imparfaite dans les zones autres que les zones de nom) s'impose comme le meilleur traitement peu importe la façon dont on mesure les résultats de l'expérience. En ce qui concerne le meilleur traitement pour l'attribution de poids pour les zones de nom, le choix doit se faire entre l'une ou l'autre des méthodes déterministes sur la fréquence. Lorsque l'une ou l'autre des méthodes déterministes est utilisée, l'usage de la méthode de comparaison de chaînes de caractères de Winkler est recommandé; lorsque c'est la méthode basée sur la fréquence qui est utilisée, il n'est pas sûr que l'on doive recourir à un comparateur de chaînes de caractères pour les noms.

Il est difficile de trancher entre les poids de "Fellegi-Sunter" pour les zones autres que les zones de nom et les poids déterminés de façon ponctuelle, mais les analyses antérieures laissent supposer que l'effet est négligeable d'un côté comme de l'autre. On peut faire les mêmes commentaires quant à l'utilisation de pré-noms normalisés ou non et l'utilisation de quatre ou sept chiffres du numéro de téléphone.

Si l'on tient compte du fait qu'il n'existe pas de combinaison de traitements qui soit supérieure à toutes les autres sur tous les plans, on peut examiner le rendement de différentes combinaisons dans une région particulière (par exemple, une région où le taux d'erreur d'appariement se situe autour de 0.001). Or, si nous regardons les meilleures combinaisons qui existent pour une région où 70 à 80% des enregistrements du fichier de l'EP appartiennent à des concordances désignées (c'est-à-dire que l'on réduit l'analyse à cinq seuils d'exclusion), aucune combinaison de traitements possible est (2,3,2,3,1,2,1,2), c'est-à-dire, attribution de poids d'une valeur de ± 4 pour les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance imparfaite dans les zones de nom, estimation de poids au moyen de l'algorithme de Fellegi-Sunter pour les zones autres que les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance imparfaite dans les zones de nom, estimation de poids au moyen de l'algorithme de Fellegi-Sunter pour les zones autres que les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance corrélée, non-inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement, et utilisation des sept chiffres du numéro de téléphone.

chaînes de caractères pour les zones de nom). Nous donnons ci-dessous dans le tableau 4.4 les résultats moyens pour toutes les combinaisons des traitements faisant intervenir les facteurs A et B.

Tableau 4.4
Rendement moyen des combinaisons des
traitements A et B

A	B	Taux d'erreur d'appariement	Arc sinus ($\sqrt{\text{fmr}}$)
1	1	0.0192	0.120
1	2	0.0140	0.099
1	3	0.0143	0.100
2	1	0.0170	0.110
2	2	0.0120	0.087
2	3	0.0123	0.089
3	1	0.0177	0.113
3	2	0.0118	0.084
3	3	0.0119	0.083
4	1	0.0254	0.145
4	2	0.0193	0.123
4	3	0.0189	0.122
5	1	0.0109	0.079
5	2	0.0109	0.078
5	3	0.0109	0.078

Par ce tableau nous voyons que lorsqu'on utilise des poids basés sur la fréquence pour le nom ($A = 5$), la méthode de comparaison de chaînes de caractères ne fait pas de différence, mais lorsqu'on utilise des poids quelconques ou des poids de "Fellegi-Sunter", l'emploi de méthodes de comparaison de chaînes de caractères accroît sensiblement l'efficacité moyenne de l'opération d'appariement informatisé.

Voici quelques-unes des autres observations intéressantes de Belin (1991) tirées de l'analyse des effets d'interactions à deux critères les plus importants:

- 1) La méthode de Winkler pour le traitement des cas de concordance imparfaite dans les zones autres que les zones de nom (c.-à-d., $D = 3$), qui est en moyenne le meilleur traitement en ce qui concerne le facteur D, est plus qu'utile en règle générale lorsque l'état matrimonial et le lien avec le chef de ménage sont inclus dans les variables d'appariement (c.-à-d., $G = 2$), même si ces variables ne sont associées d'aucune manière au traitement des cas de concordance imparfaite.
- 2) Contrairement aux autres méthodes de pondération des zones de nom, qui s'accroissent très bien de l'inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement, la méthode basée sur la fréquence semble être dérangée par une telle inclusion.
- 3) L'attribution de poids d'une valeur de ± 6 pour les zones de nom donne de meilleurs résultats en moyenne lorsqu'elle est combinée à la méthode de pondération ponctuelle pour les zones autres que les zones de nom; l'attribution de poids d'une valeur de ± 4 et de ± 2 pour les zones de nom donne de meilleurs résultats lorsque combinée à la méthode de pondération fondée sur l'algorithme de Fellegi-Sunter pour les zones autres que les zones de nom.

Belin (1991) construit un ensemble de contrastes orthogonaux complémentaires à partir des résultats expérimentaux. Les plus importants contrastes d'effets majeurs parmi ceux définis préalablement par Belin (1991) sont ceux entre les poids basés sur la fréquence pour les noms ($A = 5$) et les poids pour zones de nom de Fellegi-Sunter ($A = 4$), entre les comparateurs de chaînes de caractères de Winkler pour les zones autres que les zones de nom ($D = 3$) et les comparateurs de chaînes de caractères correspondants de Jaro ($D = 2$), entre un comparateur de chaînes de caractères quelconque pour les zones de nom pour les zones de nom ($B = 1$), entre un comparateur de chaînes de caractères quelconque pour les zones autres que les zones de nom ($D = 2$ ou 3) et aucun comparateur de chaînes de caractères pour ces mêmes zones ($D = 1$), et entre corriger des poids pour les cas de concordance corrélée ($F = 2$) et ne pas en corriger ($F = 1$).

4.4 Interactions à deux critères des traitements

Le plus important contraste d'interactions à deux critères des traitements parmi ceux analysés par Belin (1991) est l'effet $F \times G$, c'est-à-dire l'interaction de la correction de poids pour les cas de concordance corrélée (des prénoms et des seconds prénoms ou des prénoms, des sexes et des âges) et de l'inclusion (ou de la non-inclusion) de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement. Ce contraste est statistiquement significatif selon n'importe laquelle des méthodes utilisées dans Belin (1991) pour estimer un niveau de bruit sous-jacent. Dans le tableau 4.3 ci-dessous, nous donnons la valeur moyenne des variables de résultat pour les quatre combinaisons des traitements.

Tableau 4.3
Rendement moyen de combinaisons
des traitements F et G

F	G	Taux d'erreur d'appariement	Arc sinus ($\sqrt{\text{fmr}}$)
1	1	0.0182	0.116
1	2	0.0143	0.097
2	1	0.0128	0.091
2	2	0.0151	0.100

Les données de ce tableau laissent supposer que la correction de poids pour les cas de concordance corrélée (niveau 2 du facteur F) est très utile lorsque l'état matrimonial et le lien avec le chef de ménage ne sont pas inclus dans les variables d'appariement (niveau 1 du facteur G) mais qu'elle ne change rien, en moyenne, lorsque ces deux mêmes variables sont comprises dans les variables d'appariement. Le fait que nous soyons capables de constater ces effets souligne l'importance de poursuivre les études empiriques dans un cadre expérimental.

Les deux plus importants contrastes d'interactions à deux critères des traitements relevés par Belin (1991) après l'interaction $F \times G$ forment une partie de l'interaction $A \times B$ (qui consiste dans le choix de poids et de comparateurs de

3.7 Autres considérations touchant l'analyse des résultats expérimentaux

L'analyse des résultats expérimentaux partait de l'idée que des indices généraux de signification sont plus importants que des valeurs p précises, à plus forte raison s'il s'agit d'une expérience exploratoire. Belin (1991) précise que les méthodes destinées spécialement à évaluer le degré de signification tiré de ces données sont assez complexes parce que la région d'essai devrait, en réalité, être considérée comme un facteur aléatoire (si nous voulons pouvoir inférer sur les "effets déterminés par le traitement" à partir de l'échantillon de trois régions d'essai pour une population de régions possibles); cependant, les méthodes courantes qui utilisent l'interaction "région d'essai-traitement" comme terme d'erreur pour un traitement particulier ont peu d'efficacité à cause du faible nombre de régions d'essai. Belin (1991) se sert du graphique de Johnson-Tukey pour les ratios (Johnson et Tukey 1987), qui se rapproche sensiblement du graphique semi-normal de Daniel (1959), pour estimer les niveaux de bruit liés à l'évaluation de la signification des effets. Il n'est pas dans nos intentions d'exposer ici des conclusions formelles sur cette question.

4.1 Analyse de variance des résultats expérimentaux

Nous commençons par faire une analyse de variance des résultats de l'expérience factorielle en distinguant les effets déterminés par le traitement, les effets déterminés par la région d'essai, les effets déterminés par le seuil et les interactions des uns avec les autres, ces effets étant groupés selon le nombre de critères. Le tableau 4.1 contient un extrait de l'analyse de variance, qui montre les interactions de traitements (de deux à cinq critères) et les termes d'erreur correspondants.

On calcule la statistique F en divisant la moyenne des carrés pour l'effet donné par la moyenne des carrés pour l'interaction de l'effet et de la région d'essai. Par exemple, la statistique F pour les interactions à trois critères des traitements est calculée à l'aide du rapport $0,0120/0,00470 = 2,551$, le dénominateur étant le chiffre qui figure sur la ligne des interactions à quatre critères du traitement et de la région d'essai.

Tableau 4.1
Extrait de l'analyse de variance des résultats de l'expérience factorielle
(effets groupés selon le nombre de critères)

Source	df	Sommes des carrés	Moyennes des carrés	F
Effets majeurs – région d'essai	2	35,195	17,598	
Effets majeurs – traitement	13	30,917	2,378	10,570
Effets majeurs – seuil d'exclusion	12	147,515	12,293	7,548
Interactions à deux critères du traitement et de la région d'essai	26	5,850	0,225	
Interactions à deux critères des traitements	24	39,089	1,629	
Interactions à deux critères du traitement et du seuil d'exclusion	70	6,992	0,100	4,041
Interactions à trois critères du traitement et de la région d'essai	140	3,461	0,0247	
Interactions à trois critères des traitements	312	0,794	0,0025	
Interactions à trois critères du traitement et du seuil d'exclusion	206	2,472	0,0120	2,551
Interactions à quatre critères du traitement et de la région d'essai	412	1,938	0,00470	
Interactions à quatre critères des traitements	1,680	0,568	0,00034	
Interactions à quatre critères du traitement et du seuil d'exclusion	365	0,747	0,00205	2,365
Interactions à cinq critères du traitement et de la région d'essai	730	0,632	0,00087	
Interactions à cinq critères des traitements	4,944	0,226	0,00046	
Total	56,159	279,169		

3.3.4 Traitement relatif à la correction de poids composites pour les cas de concordance corrélée

On doit aussi à Winkler la méthode qui sert à corriger les poids composites de manière à tenir compte des cas possibles de concordance corrélée; cette méthode est décrite dans Winkler et Thibaudau (1992). Des études de Kelley (1986) et de Thibaudau (1989) révèlent que la concordance qui peut être observée pour des zones qui font l'objet de l'opération d'appariement entre les fichiers du recensement et ceux de l'EP n'est pas du tout indépendante d'une zone à l'autre. Les études donnent à croire, en particulier, que la concordance de prénoms est corrélée avec la concordance de seconds prénoms et que les concordances de prénoms, d'âges et de sexes sont corrélées l'une avec l'autre. Par suite de ces observations, on en est venu à appliquer des règles de modification du poids composite dans certaines circonstances (par exemple, s'il y a non-concordance des prénoms, des âges et des sexes simultanément, on soustrait du poids composite une valeur élevée). La méthode utilisée actuellement pour corriger les poids composites est une méthode entièrement *adaptée aux circonstances*; l'étude des méthodes qui tiennent compte de la concordance corrélée semble encore en être au premier stade.

3.4 Fichiers de données utilisés dans l'expérience

Comme nous l'avons mentionné plus haut, les trois régions d'essai de la répétition générale et de l'enquête postcensitaire de 1988 ont permis de constituer des fichiers de données distincts qui ont pu faire l'objet de notre analyse. Le fichier de l'EP de St-Louis contenait 12,072 enregistrements, celui du centre est du Missouri en contenait 6,581 et celui de l'Etat du Washington, 2,782. Nous avons aussi souligné que, pour les besoins de l'analyse, nous considérons comme juste le classement opéré par les commits qui avaient examiné ces fichiers. D'autres recensements d'essai ont eu lieu durant les années 1980; si nous n'avons pas tenu compte des données de ces recensements dans notre expérience, c'est surtout parce que la préparation d'un ensemble de données en vue des analyses effectuées ici exige beaucoup de temps système.

3.5 Variable de résultat

La principale variable de résultat envisagée dans cette expérience était une transformation du taux d'erreur d'appariement. Le taux d'erreur d'appariement est défini comme le quotient du nombre de fausses concordances par le nombre de concordances désignées et est une mesure de rendement couramment utilisée dans les ouvrages sur le couplage d'enregistrements (par ex., Fellegi et Sunter (1969) tentent d'obtenir des résultats qui respectent un critère fixé par l'opérateur; ce critère correspond à un taux d'erreur d'appariement). Afin de stabiliser la variance des résultats, nous utilisons l'arc sinus de la racine carrée du taux d'erreur d'appariement comme variable de résultat dans notre analyse.

3.6 Choix du poids limite comme facteur de groupage

Il est évident que dans le couplage d'enregistrements, le taux d'erreur d'appariement est susceptible de dépendre largement du choix d'un point de démarcation entre les concordances désignées et les non-concordances désignées. C'est pourquoi on introduit un facteur de groupage (facteur J) qui permet de déterminer des seuils de manière à faciliter la comparaison d'autres traitements relatifs au couplage d'enregistrements. Afin de pouvoir établir des comparaisons entre des régions d'essai qui n'ont pas le même nombre d'enregistrements, le seuil d'exclusion est exprimé en fonction de la proportion du fichier de l'EP ayant pu être couplée.

Comme les poids de couplage d'enregistrements sont des valeurs discrètes, il peut exister des liens entre les poids attribués à des paires d'enregistrements qui se trouvent dans le voisinage du seuil d'exclusion. Dans un fichier de 10,000 enregistrements par exemple, on peut en compter 40 qui ont un poids W (dont 10 peuvent appartenir à de fausses concordances), 7,980 qui ont un poids supérieur à W (dont 3 peuvent appartenir à de fausses concordances) et 1,980 qui ont un poids inférieur à W . Si, selon le facteur J, la proportion des enregistrements du fichier de l'EP pouvant être couplés doit être de 80%, il se peut que la manière de calculer le taux d'erreur d'appariement ne s'impose pas à l'évidence puisque 40 enregistrements qui ont le même poids se trouvent dans le voisinage du point où devrait être fixé le seuil d'exclusion. Dans de telles circonstances, on calcule le taux d'erreur d'appariement au moyen de la relation suivante:

$$fmr = \frac{f_{abv} + \frac{f_{bdy}}{J} \times (n^{cut} - n^{abv})}{n^{cut}},$$

où fmr désigne le taux d'erreur d'appariement, f_{abv} est le nombre de fausses concordances et n^{abv} , le nombre de concordances désignées dont le poids est supérieur au poids limite, f_{bdy} est le nombre de fausses concordances et n^{bdy} , le nombre de concordances désignées dont le poids est égal au poids limite; enfin, n^{cut} est le nombre de couples. S'il fallait calculer le taux d'erreur d'appariement en déterminant de façon aléatoire le nombre d'enregistrements limites voulu pour respecter le seuil d'exclusion, la formule ci-dessus permettrait de connaître le taux d'erreur d'appariement prévu pour plusieurs itérations de cette procédure; par conséquent, la logique sous-jacente à cette définition est claire.

Dans l'exemple ci-dessus, le quart des cas limites sont de fausses concordances, et il faut vingt enregistrements de plus pour respecter le critère de 80%. De fait, cinq fausses concordances sont ajoutées aux trois qui existent déjà parmi les enregistrements qui forment les paires dont le poids est supérieur au poids limite, ce qui donne un taux d'erreur d'appariement de $(3 + 0.25(40 - 20))/8,000 = 8/8,000 = 0.001$.

Dans le cas de certaines zones numériques (numéro de téléphone, par exemple), il serait plus raisonnable d'avoir une méthode de comparaison conçue en fonction des erreurs typographiques mineures qu'une méthode fondée sur l'écart. En revanche, pour des variables comme l'année de naissance ou l'âge, il est difficile de dire s'il faut se concentrer sur les erreurs typographiques (aucun cas une méthode de comparaison de chaînes de caractères serait la solution la plus appropriée), les erreurs de déclaration (aucun cas une méthode fondée sur l'écart serait la solution la mieux indiquée), ou d'autres types d'erreur, comme l'arrondissement de l'âge déclaré au multiple de cinq le plus près, (aucun cas ni l'une ni l'autre des méthodes de comparaison mentionnées ci-dessus ne conviendrait). C'est pourquoi nous poursuivons nos analyses empiriques dans le but d'approfondir ces questions.

3.3.3 Traitements relatifs au choix des variables d'appariement

Nous avons mentionné plus tôt que le Censur Bureau avait élaboré une méthode pour convertir les diminutifs dans une forme "normalisée". Le logiciel élaboré par Palletz (1989) applique le programme de normalisation des noms.

Le traitement qui n'inclut pas l'état matrimonial et le lien avec le chef de ménage dans les variables d'appariement permet d'évaluer l'importance de deux variables démographiques de base pour la qualité de l'appariement. Chernoff (1980) élabore une théorie relative à l'information que "renferme" une variable d'appariement et il montre qu'une variable qui est enregistrée incorrectement, ne serait-ce que très peu de fois, peut "perdre" une partie appréciable de l'information qu'elle contient pour l'appariement (par exemple, l'information de Kullback-Leibler rattachée à une variable binaire qui a été enregistrée incorrectement dans trois pour-cent des cas n'équivaut qu'à environ la moitié de l'information rattachée à une variable binaire qui a enregistré correctement à tout coup). Compte tenu de ce que le lien avec le chef de ménage pourra avoir changé entre le recensement et l'EP si la personne identifiée à l'origine comme le chef de ménage n'est plus la même, et de ce que l'état matrimonial de certaines personnes changera dans l'intervalle, on ne peut dire d'avance quelle quantité d'information ces variables renferment-elles pour l'appariement. En revanche, il est difficile de penser que l'utilisation de variables d'appariement additionnelles serait nuisible; par conséquent, ce traitement peut servir de norme pour l'évaluation de l'importance pratique de certains des autres traitements.

L'utilisation de quatre ou sept chiffres du numéro de téléphone comme variable d'appariement est explicite. Ce traitement est justifié par le fait que l'élaboration de l'une des méthodes de comparaison de chaînes de caractères de Winkler (métriques linéaires progressives) reposait sur l'analyse des quatre derniers chiffres du numéro de téléphone comme variable d'appariement.

de pondération, où la fréquence relative des noms dans les fichiers disponibles est prise en considération. Cette méthode consiste à attribuer un poids de concordance plus grand pour des noms comme Abramowicz, qui peuvent être relativement rares, que pour des noms comme Smith, qui peuvent être très courants. Il se pourrait, évidemment, que Abramowicz soit plus courant que Smith dans une région particulière; dans ce cas, on attribuerait un poids de concordance plus grand pour le nom Smith. L'idée de tenir compte de fréquences marginales tirées des fichiers courants a été lancée par Newcombe *et al.* (1959) et est évoquée depuis par de nombreux auteurs, dont Fellegi et Sunter (1969). (Par conséquent, la distinction qui est faite entre l'"algorithme de Fellegi-Sunter" et la "pondération fondée sur la fréquence" sert en réalité à distinguer deux méthodes de calcul de poids analysées l'une et l'autre par Fellegi et Sunter.) On trouvera dans Winkler et Thibaudreau (1992) des détails sur la mise en application de la méthode de pondération fondée sur la fréquence dans le programme du Censur Bureau.

3.3.2 Traitement des cas de concordance imparfaite

Les facteurs B et D ont rapport au traitement des cas de concordance imparfaite (c.-à-d. des cas où deux zones d'information se rapprochent sensiblement l'une de l'autre sans être parfaitement identiques). Plusieurs techniques ont été proposées pour le traitement des cas de concordance imparfaite; ces techniques sont souvent aussi nombreuses que les théories qui cherchent à expliquer l'absence de concordance parfaite.

Le comparateur de chaînes de caractères de Jaro sert à mesurer le degré de concordance de deux zones à caractères multiples; la métrique qui définit le degré de similitude est fonction de la longueur des zones de caractères des deux fichiers, du nombre de caractères communs aux deux zones et du nombre de caractères qui peuvent être transposés d'une zone à l'autre. Le poids attribué pour une concordance partielle se situe entre celui attribué pour une concordance de zones et celui attribué pour une non-concordance de zones et est une fonction linéaire du comparateur de chaînes de caractères, ce qui nécessite deux paramètres de taux et deux seuils fournis par l'utilisateur au point de changement de la pente.

La méthode proportionnelle de Jaro permet d'attribuer un poids dont la valeur se situe entre le poids de concordance et le poids de non-concordance, en se fondant sur la valeur absolue de la différence entre deux zones numériques. Comme pour les techniques précédentes, le poids de concordance partielle se trouve être une fonction linéaire de la valeur absolue de la différence.

de même que des méthodes fondées sur des estimations de paramètres de modèles probabilistes explicites. L'étude de méthodes de pondération ponctuelles nous donne l'occasion de jauger l'importance d'introduire des méthodes plus

L'expression "algorithme de Fellegi-Sunter" désigne la méthode exposée dans l'article de Fellegi et Sunter (1969), qui repose sur un modèle probabiliste qui intègre de l'information sur des schémas de concordance et de non-concordance pour des paires d'enregistrements. Le modèle suppose que les probabilités de concordance de zones d'information, étant donné qu'une paire représente une vraie concordance ou bien une fausse concordance, sont indépendantes d'une zone à l'autre. Dans leur article, Fellegi et Sunter montrent que ce modèle implique certaines propriétés d'optimalité pour le genre de méthode de pondération qu'utilisent Newcombe *et al.* (1959), selon laquelle on calcule les poids relatifs aux zones d'information par le logarithme du rapport de la probabilité de concordance étant donné une concordance vraie à la probabilité de concordance étant donné une fausse concordance, et selon laquelle on calcule les poids composites en faisant la somme des poids de chaque zone.

3.3 Aspects particuliers des traitements expérimentaux

3.3.1 Traitements relatifs à l'attribution de poids pour les cas de concordance ou de non-concordance dans des zones d'information

Afin de préciser l'expérience, nous décrivons plus en détail chacun des facteurs expérimentaux. Les facteurs A et C ont trait à l'attribution de poids de concordance ou de non-concordance pour diverses variables d'appariement. Les méthodes de pondération utilisées dans les circons-tances comprennent des méthodes tout à fait ponctuelles

Symbole	Description du facteur	Nombre de niveaux du facteur	Description des niveaux
D	Attribution de poids pour les cas de concordance imparfaite dans les zones autres que les zones de nom.	3	<p>1. Attribution d'un poids de non-concordance pour toute différence entre des zones autres que les zones de nom.</p> <p>2. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de numéros de rue, de numéros de téléphone, d'âges, en utilisant le comparateur de chaînes de caractères de Jaro.</p> <p>3. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de noms de rue, en utilisant le comparateur de chaînes de caractères de Jaro; s'il y a quasi-concordance d'âges, en utilisant la métrique "au prorata de l'écart absolu" de Jaro; s'il y a quasi-concordance de numéros civiques ou de numéros de téléphone, en utilisant le comparateur linéaire progressif de chaînes de caractères de Winkler.</p>
E	Utilisation du prénom introduit au clavier ou de son équivalent normalisé.	2	<p>1. Utilisation du prénom introduit au clavier dans chaque fichier pour la comparaison de prénom.</p> <p>2. Utilisation du prénom produit par le logiciel de normalisation des noms (Paltz 1989).</p>
F	Correction de poids pour les cas de concordance corrigée.	2	<p>1. Ne pas corriger les poids composites pour les cas possibles de concordance corrigée.</p> <p>2. Corriger les poids composites pour les cas possibles de concordance corrigée des prénom et des seconds prénom et de concordance corrigée des prénom, des sexes et des âges.</p>
G	Inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement.	2	<p>1. Ne pas compter l'état matrimonial et le lien avec le chef de ménage parmi les variables d'appariement.</p> <p>2. Compter l'état matrimonial et le lien avec le chef de ménage parmi les variables d'appariement.</p>
H	Utilisation de quatre ou sept chiffres pour le numéro de téléphone.	2	<p>1. Utiliser seulement les quatre derniers chiffres du numéro de téléphone comme variable d'appariement.</p> <p>2. Utiliser les sept chiffres du numéro de téléphone.</p>
I	Région d'essai du recensement ou de l'enquête postcensitaire.	3	<p>1. Est de l'Etat du Washington.</p> <p>2. Columbia, Missouri.</p> <p>3. St-Louis, Missouri.</p>
J	Proportion du fichier de l'EP ayant pu être couplée.	13	<p>1. - 13. Le nombre d'enregistrements ayant pu être couplés équivalait à 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, 90% du nombre total d'enregistrements de l'EP dans la région d'essai donnée.</p>

le plus semblables possible en l'absence d'effets déterminés par le traitement) avec répétition dans trois régions d'essai par un plan factoriel $2^5 \times 3^3 \times 5 \times 13$. La variable de résultat de l'expérience, décrite plus loin dans la section 3.5, consistait en une transformation du taux d'erreur d'appariement, cette transformation étant destinée à stabiliser la variance des résultats. Les facteurs utilisés dans l'expérience peuvent être décrits comme suit:

Symbole	Description du facteur	Nombre de niveaux du facteur	Description des niveaux
---------	------------------------	------------------------------	-------------------------

A	Attribution de poids pour les zones de nom.	5	1. Attribution de poids d'une valeur de ± 2 selon qu'il y a concordance ou non concordance pour le prénom, le nom de famille.
---	---	---	---

			2. Attribution de poids d'une valeur de ± 4 selon qu'il y a concordance ou non concordance pour le prénom, le nom de famille.
--	--	--	---

			3. Attribution de poids d'une valeur de ± 6 selon qu'il y a concordance ou non concordance pour le prénom, le nom de famille.
--	--	--	---

			4. Attribution de poids en fonction des estimations de la probabilité de concordance du prénom, du nom de famille, calculées selon l'algorithme de Fellegi-Sunter (voir Winkler et Thibaudreau 1992).
--	--	--	---

			5. Utilisation de la pondération basée sur la fréquence pour le prénom, le nom de famille (voir Winkler et Thibaudreau 1992).
--	--	--	---

B	Attribution de poids pour les cas de concordance imparfaite dans les zones de nom.	3	1. Attribution d'un poids de non-concordance pour toute différence entre prénoms, entre noms de famille.
---	--	---	--

			2. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de prénoms, de noms de famille, en utilisant le comparateur de chaînes de caractères de Jaro (Jaro 1989; Winkler 1991).
--	--	--	--

			3. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de prénoms, de noms de famille, en utilisant la métrique linéaire progressive décrite dans Winkler (1991).
--	--	--	---

C	Attribution de poids pour les zones autres que les zones de nom.	2	1. Attribution de poids d'une valeur de ± 2 selon qu'il y a concordance ou non-concordance pour l'âge, le numéro de téléphone ou l'adresse, et attribution de poids d'une valeur de ± 1 selon qu'il y a concordance ou non-concordance pour le sexe, l'origine raciale, l'état matrimonial, le lien avec le chef de ménage ou le second prénom.
---	--	---	---

			2. Attribution de poids en fonction des estimations de la probabilité de concordance calculées selon l'algorithme de Fellegi-Sunter.
--	--	--	--

3.2 Expérience factorielle réalisée avec des données du recensement et de l'enquête postcensitaire

On a réalisé une étude à l'aide de données provenant de chacune des trois régions d'essai (St-Louis, Missouri; du Washington) de la répétition générale et de l'EP de 1988. Ces ensembles de données ont été apparées par ordinateur, puis examinées par des commis. Pour les besoins des analyses qui vont suivre, nous tenons pour acquis que le classement opéré par les commis en ce qui a trait aux concordances (vraie concordance/fausse concordance) est juste. Par conséquent, malgré que les analyses qui vont suivre ne pourront jamais être plus précises que celles des commis, ces fichiers de données nous offrent une excellente occasion d'étudier le couplage d'enregistrements.

On trouvera une description des méthodes particulières utilisées pour coupler les enregistrements du recensement à ceux de l'EP dans Jaro (1989), Winkler (1991) et Winkler et Thibaudau (1992). Selon la méthode actuellement en usage, l'utilisateur jouit de plusieurs possibilités par rapport à tous les facteurs qui ont été énumérés dans la section 3.1, sauf en ce qui concerne le choix de l'algorithme servant à rapprocher des éléments susceptibles de former la même paire (on utilise un algorithme de "rapprochement à somme linéaire", voir Jaro 1989).

Les variables qui peuvent servir à l'appariement des enregistrements du recensement avec ceux de l'EP sont le nom, l'adresse, l'âge, l'origine raciale, le sexe, le numéro de téléphone, l'état matrimonial et le lien avec le chef de ménage. En pratique, le nom est habituellement décomposé en trois éléments: prénom, nom de famille et second prénom, chacun d'eux servant de variable d'appariement. On utilise habituellement un programme de prétraitement pour décomposer l'adresse en ses éléments: numéro civique, rue, numéro d'appartement, numéro de route rurale et numéro de case postale (Laplant 1989). Il arrive qu'en raison d'"irrégularités" dans la zone "adresse", causées probablement par une erreur de saisie ou une erreur d'inscription de la part d'un intervieweur, on ne soit pas capable de décomposer une adresse en ses éléments; dans ces circonstances, la zone "adresse" intégrale (appelée "adresse globale") sert de variable d'appariement. On peut aussi se servir d'un programme de prétraitement pour convertir les diminutifs dans leur forme "normalisée" en utilisant une banque de noms, y compris les variantes les plus courantes (Paletz 1989). Il existe diverses méthodes permettant d'attribuer des poids en fonction d'une quasi-concordance de variables, et il existe aussi une méthode qui permet d'accroître ou de réduire (par une simple opération d'addition ou de soustraction) le poids composite associé à une paire d'enregistrements lorsque certaines combinaisons de zones concordent ou ne concordent pas entre elles (Winkler 1991).

L'expérience a consisté en huit facteurs de "traitement" et en un facteur de "groupe" (ce terme signifie ici, par rapport à la planification d'expériences, que l'on forme un groupe d'unités qui sont censées produire des résultats

lesquels on observe un écart entre les données du recensement et les données de l'EP. La distinction entre les concordances possibles et les non-concordances a trait uniquement aux méthodes qu'appliquent les commis lorsqu'ils examinent ces cas (Childers 1989; Donoghue 1990). Dans le traitement des données du recensement et de l'EP de 1990, l'opérateur qui voit à l'exécution du programme d'appariement détermine manuellement des poids limites qui permettent d'identifier les concordances, les concordances possibles et les non-concordances après avoir exploré des ensembles d'éléments susceptibles de former la même paire et affectés de poids dont la valeur se situe dans un intervalle donné. Une nouvelle méthode proposée par Belin et Rubin (1991) permet de déterminer automatiquement les poids limites.

3.1 Facteurs qui influent sur les résultats des méthodes de couplage d'enregistrements

L'efficacité d'une méthode de couplage d'enregistrements tient à un certain nombre de facteurs, dont voici la liste:

- (1) le choix des variables d'appariement;
- (2) le choix des variables de groupe;
- (3) l'attribution de poids de concordance ou de non-concordance pour diverses variables d'appariement;
- (4) le traitement des cas de concordance imparfaite entre des variables d'appariement;
- (5) le traitement de données manquantes dans l'un ou l'autre des enregistrements formant une paire ou dans les deux enregistrements;
- (6) l'algorithme servant à rapprocher des éléments susceptibles de former la même paire;
- (7) le choix d'un poids limite au-delà duquel des paires d'enregistrements seront reconnues comme des concordances;
- (8) le lieu ou le milieu d'où proviennent les données.

Parmi tous ces facteurs, seul le huitième représente une source de variation sur laquelle l'opérateur qui voit à l'exécution du programme d'appariement n'a aucun contrôle. Comme nous l'avons mentionné plus tôt, deux questions en particulier méritent d'être approfondies dans l'expérience factorielle. Primo, si on parvient à déterminer les principales sources de variation dans le couplage d'enregistrements, on sera en mesure d'orienter les recherches en matière de couplage pour l'avenir et de mieux comprendre le processus par lequel des erreurs sont introduites dans les méthodes de couplage. Secundo, il serait intéressant de déterminer la combinaison de facteurs qui permet d'obtenir le plus grand nombre de concordances avec le taux d'erreur le plus faible possible puisque dans la pratique, l'utilisateur a généralement le choix entre de nombreuses possibilités pour chaque facteur que nous venons de décrire.

une représentation graphique des principales étapes du couplage des enrégistrement du recensement et de l'EP.

La base du recensement est constituée de listes d'adresses d'unités de logement. Ces listes sont formées à l'aide de diverses méthodes, celles-ci variant généralement selon qu'il s'agit d'une région urbaine ou d'une région rurale. En ce qui concerne les régions urbaines et suburbaines, on envoie les questionnaires du recensement par la poste en espérant que les ménages les retourneront dûment remplis; pour ce qui est des autres régions, des recenseurs se rendent à domicile. Lorsqu'un ménage ne retourne pas le questionnaire qu'il a reçu par la poste, un recenseur se rend au domicile de ce ménage. Les données sont versées dans les fichiers informatiques du Census Bureau par une combinaison de techniques d'exploration informatisée et d'opérations de saisie clavier. Pour un aperçu de la méthodologie du recensement, voir Citro et Cohen (1985); on trouvera une description détaillée de diverses opérations du recensement dans le recueil du Bureau of the Census intitulé '1990 Decennial Census Information Memorandum Series' (Bureau of the Census 1988-1991).

Dans des enquêtes postcensitaires comme celle qui a été réalisée en 1990 (et dans des recensements d'essai comme ceux qui ont mené à l'EP de 1990), la collecte de données débute par la confection de listes d'adresses; cette opération est confiée à des recenseurs qui parcourent les quartiers. La collecte proprement dite se fait uniquement par interview sur place, contrairement à la méthode 'envoi et retour par la poste' utilisée dans le recensement pour les régions urbaines et suburbaines. Les données pertinentes sont versées dans les fichiers informatiques uniquement par saisie clavier. Hogan (1992) donne un aperçu de l'EP; on trouvera une description détaillée des opérations de l'EP dans le recueil du Bureau of the Census intitulé 'STSD Decennial Census Memorandum Series' (Bureau of the Census 1987-1991).

Le prétraitement des données est un sujet qui est peu souvent abordé dans les ouvrages sur le couplage d'enregistrements, même si cette étape est celle où existe la possibilité de tirer toutes les informations voulues des données d'enquête comme la possibilité de rejeter inconsiderément des informations utiles. Winkler (1985a, 1985b) présente des méthodes particulières qui ont pour propriété de faciliter la distinction entre de vraies concordances et de fausses concordances, et Jabine et Schuren (1986) de même que Newcombe (1988) proposent de grandes lignes directrices à ce sujet. Dans l'opération d'appariement des enrégistrement du recensement à ceux de l'EP, le prétraitement des données consiste à coder les variables démographiques selon des normes, à décomposer la zone d'adresse en ses éléments préalablement définis (numéro civique, rue, numéro d'appartement, numéro de route rurale, numéro de case postale) (Laplant 1989), et à 'normaliser' les prénoms d'individus en confrontant le prénom introduit au clavier à une liste de diminutifs puis en convertissant les diminutifs observés dans leur forme au long (Paletz 1989).

Le couplage des enrégistrement du recensement à ceux de l'EP est un procédé fondé sur la pondération. Pour déterminer une méthode de pondération, il faut envisager des règles de modèle et des règles *ad hoc* pour l'attribution de poids pour la concordance ou la non-concordance de zones d'information, pour la concordance imparfaite de zones d'information, pour des enrégistrement incomplets et pour la concordance ou la non-concordance de certaines combinaisons de variables.

La définition d'éléments susceptibles de former la même paire dans l'opération d'appariement des enrégistrement du recensement à ceux de l'EP reflète certaines contraintes auxquelles est assujéti le processus d'appariement. Premièrement, à cause de contraintes de temps et de limites financières, il est peu réaliste de comparer chacun des enrégistrement d'un fichier à chaque enrégistrement d'un autre fichier. Par conséquent, on limite la comparaison à des paires d'enregistrements qui répondent à certains critères minimum (par ex., les deux enrégistrement comparés doivent provenir du même îlot de recensement et la première lettre du nom de famille doit être la même dans les deux cas). Le sous-ensemble d'enregistrements ainsi constitué est appelé 'groupe' et les variables par rapport auxquelles on doit observer une concordance de deux enrégistrement sont appelées 'variables de groupe' (Jaro 1989).

Une autre contrainte à laquelle est assujéti le processus d'appariement des enrégistrement du recensement à ceux de l'EP est celle voulant qu'un enrégistrement donné d'un fichier ne peut être couplé à plus d'un enrégistrement d'un autre fichier. La méthode utilisée pour rapprocher des éléments susceptibles de former la même paire s'inspire des techniques de recherche opérationnelle appliquées au problème de transport (Jaro 1989). L'algorithme rapproche des éléments susceptibles de former une concordance afin de maximiser la somme des poids composites qui se rattachent à chacune des paires d'enregistrements d'un groupe défini par les variables de groupe, sous réserve qu'un enrégistrement d'un fichier ne peut être couplé à plus d'un enrégistrement d'un autre fichier. Supposons par exemple que dans un groupe particulier, l'enregistrement A du fichier 1 a un poids de concordance plus élevé pour l'enregistrement B du fichier 2 que pour tout autre enrégistrement du même fichier. L'algorithme de rapprochement peut néanmoins coupler l'enregistrement A à un autre enrégistrement que B, disons C, et coupler l'enregistrement B à un autre enrégistrement que A, disons D, si la somme des poids de concordance pour (A, C) et (D, B) est plus élevée que celle pour toute autre combinaison de couples.

Selon la méthode utilisée actuellement pour appairer les enrégistrement du recensement à ceux de l'EP, l'ordinaire a trois possibilités d'action: reconnaître une paire d'enregistrements comme une concordance, reconnaître une paire d'enregistrements comme une non-concordance possible ou reconnaître une paire d'enregistrements comme une non-concordance. Tous les cas de non-concordance et de concordance possible sont soumis à l'attention de commis pour être examinés, et on tente d'organiser une interview de suivi avec les ménages pour

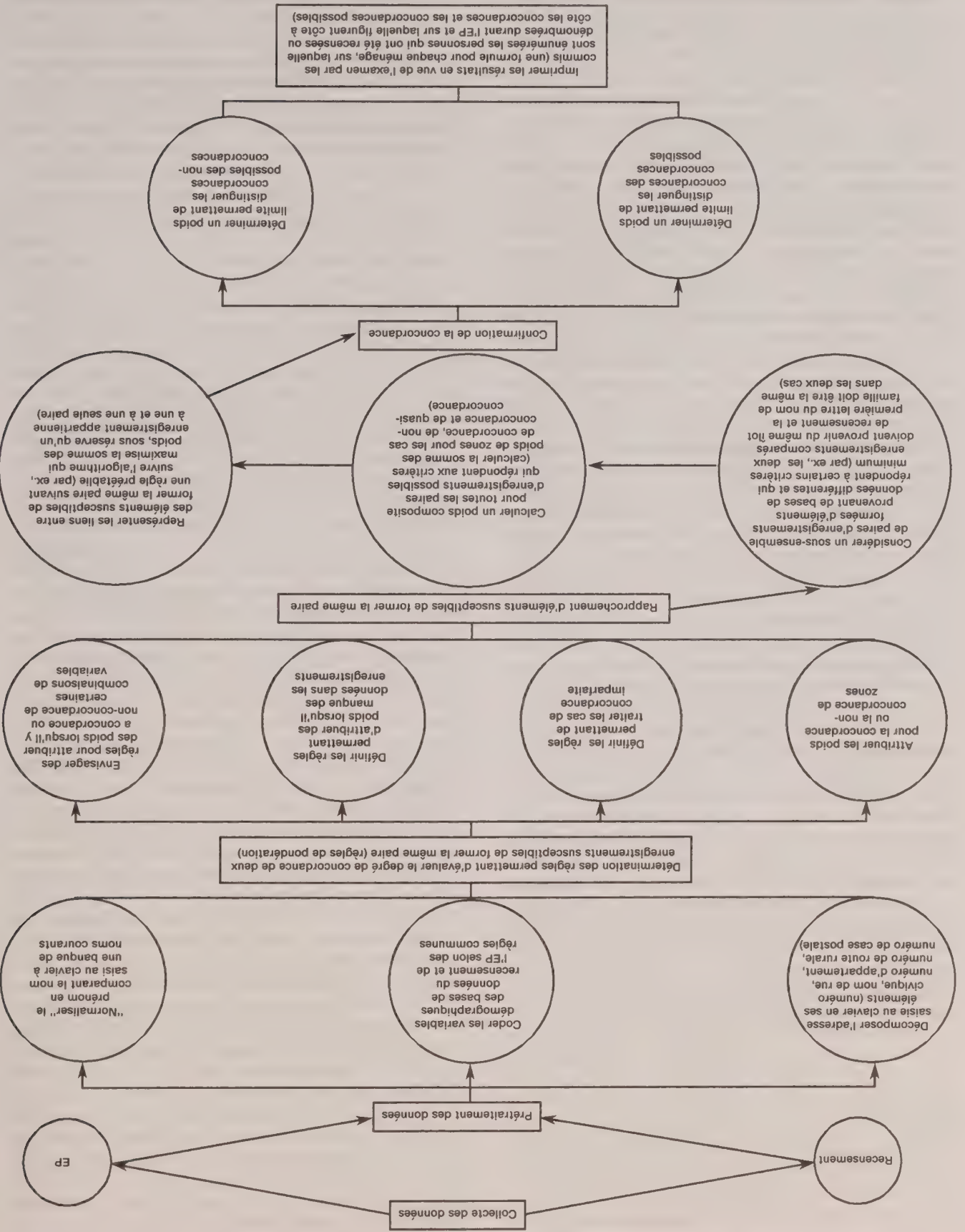


Figure 1. Etapes de la méthode de couplage des enregistrements du recensement avec ceux de l'EP

d'enregistrements qui présuppose une relation de concordance indépendante pour chacune des zones de données d'un enregistrement. Fellegi et Sunter montrent que si on utilise une méthode de pondération semblable à celle de Newcombe *et al.* avec des poids limites qui dépendent d'un certain taux d'erreur dans les concordances et d'un certain taux d'erreur dans les non-concordances, on se trouve à définir une méthode de couplage qui est optimale en ce sens qu'elle réduit au maximum la proportion d'enregistrements qui ne peuvent être classés ni dans les concordances désignées ni dans les non-concordances désignées, cela en supposant que le modèle utilisé est valide.

Par la suite, le développement des méthodes de couplage d'enregistrements s'est surtout fait à la faveur d'applications, les chercheurs mettant en pratique les théories exposées dans les ouvrages antérieurs. Parmi les applications les plus notables, mentionnons la Oxford Record Linkage Study (Acheson 1967; Goldacre 1986), l'opération d'appariement effectuée avec des enregistrements de trois sources: la Current Population Survey, la Social Security Administration et l'Internal Revenue Service (Kilss et Scheuren 1978), et la National Longitudinal Mortality Study (Rogot, Sorlie, Johnson, Glover et Treasure 1988). Les actes de colloques portant sur le couplage d'enregistrements (Kilss et Alvey 1985; Howe et Spasoiff 1986; Carpenter et Fair 1990), les recueils de symposiums annuels (Kilss et Alvey 1984a; Kilss et Alvey 1984b; Kilss et Alvey 1984c; Kilss et Alvey 1987; Kilss et Jamerson 1990), de même que les actes de colloques portant plus spécialement sur les utilisations des données administratives (Coombs et Singh 1988) décrivent de nombreuses autres applications où l'on utilise des méthodes de couplage d'enregistrements.

L'élaboration de logiciels a fourni de nouveaux outils pour poursuivre les recherches dans le domaine du couplage d'enregistrements. À Statistique Canada comme au Bureau of the Census des E.-U., on a élaboré pour des applications variées des logiciels qui apportent des améliorations aux méthodes de pondération et de groupage. Howe et Lindsay (1981) font état du "Système itératif général de chaînage d'articles" (SIGCA) de Statistique Canada; les ouvrages de Hill (1981) et de Hill et Pring-Mill (1986) portent aussi sur ce système. Du côté américain, les ouvrages de Jaro (1989), Winkler (1989) et Winkler et Thibaudeau (1992) font état du système d'appariement mis au point par la division du couplage d'enregistrements du Bureau of the Census, tandis que les deux ouvrages de Laplant (1988 et 1989) et celui de Winkler (1991) renferment de la documentation sur ce système.

Copas et Hilton (1990) présentent de nouveaux modèles qui reflètent les particularités des fichiers de données qui peuvent servir à l'élaboration d'une méthode de pondération probabiliste. Newcombe, Fair et Lalonde (1992) décrivent d'autres méthodes de couplage conçues pour tirer profit de l'information contenue dans les noms de personnes. Comme ouvrages de référence sur le couplage d'enregistrements, mentionnons l'article de synthèse de Jabiné et Scheuren (1986), le manuel de Newcombe (1988) ainsi que l'ouvrage édité par Baldwin, Acheson et Graham (1987).

2.3 Étapes d'une méthode type de couplage d'enregistrements

Les étapes habituelles d'une méthode de couplage d'enregistrements sont les suivantes: 1) collecte de données, 2) prétraitement des données, 3) définition des règles permettant d'évaluer le degré de concordance de deux éléments susceptibles de former la même paire, 4) rapprochement d'éléments susceptibles de former la même paire et 5) confirmation de la concordance. Nous employons l'expression "éléments susceptibles de former la même paire" pour décrire des couples d'enregistrements qui ont des chances de constituer la meilleure concordance possible parmi tous les éléments de leurs fichiers respectifs (voir les termes "hits" dans Rogot, Sorlie et Johnson (1986), "pairs" dans Winkler (1989) et "assigned pairs" dans Jaro (1989)). Il se peut que des éléments susceptibles de former la même paire soient confirmés dans leur concordance à la suite de l'application d'une règle de décision à l'étape (5), mais ils ne le seront pas nécessairement par la règle proprement dite.

Comme nous l'avons mentionné plus tôt, dans beaucoup de méthodes de couplage, on évalue le degré de concordance de deux éléments susceptibles de former la même paire au moyen d'une statistique globale unidimensionnelle souvent appelée "poids composite". Dans de telles méthodes, l'étape (3) mentionnée ci-dessus consisterait dans la définition de règles de pondération et l'étape (5) consisterait à fixer un poids limite au-delà duquel des paires d'enregistrements seraient reconnues comme des concordances.

Le couplage d'enregistrements peut être envisagé comme un problème décisionnel où l'ordinateur a deux possibilités d'action ou plus. En règle générale, on parle de trois possibilités d'action (par ex., reconnaître une concordance, reconnaître une non-concordance ou confier les enregistrements à un observateur pour que celui en fasse un examen plus profond, comme dans Fellegi et Sunter 1969), bien que parfois seulement deux possibilités soient envisagées (reconnaître une concordance, reconnaître une non-concordance). Il y a même des cas où le nombre de possibilités d'action atteint cinq (Tepping 1968). Le fait de postuler que la distance entre des enregistrements multidimensionnels peut être résuée par un poids composite unidimensionnel réduit l'éventail de méthodes qui peuvent servir au couplage d'enregistrements. Nous connaissons très peu d'études où l'on examine d'autres voies que celle évoquée ci-dessus, si nous faisons abstraction de la simple définition d'un ensemble déterministe de règles qui permettent de dire si une paire d'enregistrements constitue une concordance. L'étude de Smith et Newcombe (1975) est l'une des rares où, effectivement, de nouvelles voies sont envisagées. Cependant, toute cette question dépasse le cadre de notre article.

2.4 Description détaillée de la méthode servant à appairer les enregistrements du recensement à ceux de l'EP

Diverses techniques peuvent être incluses dans chacune des cinq étapes mentionnées plus haut. La figure 1 donne

d'un recensement aux enregistrements d'une enquête post-censitaire (EP) générale exécutée à la suite du recensement afin d'évaluer le taux de couverture de ce recensement. Comme autres exemples d'application de la seconde catégorie, mentionnons l'étude de Nicholl (1986) sur les erreurs de classification concernant les genres de blessures subies par les victimes d'accidents de la circulation (étude fondée sur le rapprochement de dossiers d'hôpitaux et de rapports de police), l'étude de Johnson (1991) sur le volume de travail des procureurs du ministère de la justice des E.-U. dans différents districts (étude fondée sur le couplage d'une liste de cas dressée par le ministère de la justice et d'une liste de cas dressée par les tribunaux fédéraux de première instance), ainsi que diverses études portant sur l'exactitude et le champ d'observation des fichiers de données sur la mortalité (Wentworth *et al.* 1983; Curb *et al.* 1985; Boyle et Decoufle 1990; Williams *et al.* 1992).

L'estimation du sous-dénombrement dans le recensement a été un sujet notable et, par moments, controversée en recherche statistique, surtout durant la dernière décennie. La controverse porte essentiellement sur une proposition voulant que l'on redresse les chiffres du recensement en se fondant sur des estimations du sous-dénombrement tirées d'une EP (enquête postcensitaire). Pour se documenter sur la question, le lecteur se référera aux ouvrages suivants: Erickson et Kadane (1985), Citro et Cohen (1985), Freedman et Navidi (1986), Wolter (1986), Schrim et Preston (1987), Erickson, Kadane et Tukey (1989), Cohen (1990) ainsi que les sections spéciales sur l'erreur de couverture dans le recensement des numéros de juin et décembre 1988 de cette revue. Le couplage d'enregistrements est la première étape dans l'appariement de données du recensement et de données de l'EP; cette étape est suivie d'un appariement d'enregistrements effectué par des commis, puis d'une interview de rappel menée auprès de ménages lorsqu'il semble y avoir des divergences entre les données du recensement et celles de l'EP et, enfin, d'une autre séance d'appariement effectuée par des commis. En s'appuyant sur des rapports de l'opération d'appariement et certaines hypothèses relatives à la probabilité qu'un individu figure uniquement dans les enregistrements du recensement, uniquement dans ceux de l'EP, dans les deux à la fois ou ni dans l'un ni dans l'autre, on peut estimer les taux de sous-dénombrement (ou de surdénombrement) dans le recensement.

2.2 Généralités sur la théorie du couplage d'enregistrements

Le développement de l'approche probabiliste dans la théorie du couplage d'enregistrements remonte à Newcombe, Kennedy, Axford et James (1959), qui ont élaboré une méthode de pondération dans le but de représenter la probabilité que deux enregistrements concordent l'un avec l'autre. Fellegi et Sunter (1969) approfondissent les fondements théoriques des règles de pondération courantes et font remarquer que la méthode proposée par Newcombe *et al.* équivalait à calculer un rapport de vraisemblance suivant un modèle simple pour couplage

évaluation systématique). L'idée de réaliser une expérience ou un ingénieur du contrôle de la qualité, mais il semble qu'elle ne soit pas retenue en ce qui concerne le couplage d'enregistrements, si l'on fait abstraction de la présente étude et de certains de nos ouvrages antérieurs (Belin 1989a, 1989b).

2. DOMAINES D'APPLICATION DU COUPLAGE D'ENREGISTREMENTS

2.1 Applications pour le couplage d'enregistrements

Les méthodes de couplage d'enregistrements sont utilisées dans diverses circonstances. On peut diviser les applications en deux grandes catégories: 1) celles où on cherche à faire des inférences sur les relations entre des variables provenant de grands fichiers différents et 2) celles où on s'intéresse spécialement au nombre d'individus qui sont représentés dans un fichier ou les deux fichiers (ou à une fonction de ces quantités).

Les exemples d'applications de la première catégorie abondent. En effet, citons les études où l'on rapproche les données d'enquêtes de santé et de nutrition avec les données des registres de décès afin d'étudier le rapport entre les facteurs de risque alimentaires et les décès attribuables à diverses causes (Johansen 1986), les études où l'on rapproche des données de l'enquête sur la population active avec des données sur la mortalité afin d'évaluer les effets de l'exploitation des gîtes d'uranium sur la santé (Newcombe, Smith, Howe, Mingay, Strugnell et Abbat 1983; Abbat 1986), les études où l'on rapproche des données sur la formation scolaire d'individus avec des données sur les traitements que reçoivent ces individus afin d'évaluer les avantages d'une formation collégiale (Fagerlin 1975), les études où l'on compare le revenu indiqué dans les dossiers de l'aide sociale avec celui indiqué dans les dossiers d'impôt (Kershaw et Fair 1979), et les études où l'on rapproche les dossiers d'individus exposés à des radia-tions lors de tests nucléaires et les dossiers d'une cohorte d'individus-témoins avec les registres nationaux des décès afin d'évaluer les différences de régime de mortalité entre les individus exposés et les individus-témoins (Dulberg Spasoff et Raman 1986). Dans toutes ces études, l'utilisation de méthodes de couplage d'enregistrements présente de l'intérêt surtout à cause du coût relativement peu élevé de ces méthodes et de leur rapidité d'exécution, puisqu'il est beaucoup plus long et beaucoup plus coûteux de réaliser les mêmes études avec une ou plusieurs étapes de suivi que de se servir des données existantes.

Le principal exemple que nous présentons dans cet article est représentatif des applications de la seconde catégorie, qui visent à déterminer le nombre de cas communs à deux fichiers de données. Dans l'exemple en question, on utilise une méthode de couplage d'enregistrements dans la première étape d'une opération d'appariement de grande envergure qui permet de comparer les enregistrements

Evaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle

THOMAS R. BELIN¹

RÉSUMÉ

Le couplage d'enregistrements désigne une technique algorithmique qui sert à identifier des paires d'enregistrements ayant trait au même individu dans des fichiers distincts. Dans cet article, nous étudions un modèle destiné à évaluer les sources de variation dans le couplage d'enregistrements en comparant ce procédé à une "boîte noire" qui reçoit des données d'entrée et restitue un produit (un ensemble de concordances désignées, c.-à-d. de paires dont les éléments représentent la même entité) qui a certaines caractéristiques. Nous illustrons nos propos au moyen d'une expérience factorielle dans laquelle nous servons de données du recensement et d'enquêtes postcensitaires afin d'évaluer l'influence de divers facteurs qui sont réputés pour réduire la fiabilité du procédé. Avec ce cadre expérimental, l'évaluation du couplage d'enregistrements devient un problème statistique comme les autres. L'étude permet de répondre à plusieurs questions de recherche et nous prétendons qu'il est essentiel de recourir à des méthodes expérimentales comme celle proposée ici si l'on veut mieux comprendre les sources d'erreur qui interviennent dans les techniques de couplage d'enregistrements.

MOTS CLÉS: Poids limite; taux d'erreur d'appariement; algorithme de Fellegi-Sunter; variables d'appariement; enquête postcensitaire; comparaison de chaînes de caractères; méthode de pondération.

1. ÉVALUATION DES MÉTHODES DE COUPLAGE D'ENREGLISTREMENTS

Le couplage d'enregistrements désigne une technique algorithmique qui sert à distinguer les paires d'enregistrements dont les éléments se rapportent au même individu mais proviennent de fichiers distincts. Le but de l'opération est de définir par une méthode informatique les paires d'enregistrements de fichiers différents qui devraient être désignées comme des "concordances" et celles qui devraient être désignées comme des "non-concordances" et ce, avec un taux d'erreur raisonnable. Ainsi, on évite les coûts d'un traitement manuel.

Pour définir une méthode de couplage d'enregistrements, il faut avoir une méthode qui permet de mesurer le degré de concordance d'enregistrements ainsi qu'une règle qui permet d'identifier des paires d'enregistrements comme des concordances ou des non-concordances. Dans les ouvrages portant sur le couplage d'enregistrements, on a beaucoup parlé du problème qui peut se poser lorsqu'on attribue des "poids" à des zones d'information dans un enregistrement à variables multiples pour obtenir un "poids composite" qui résume le degré de concordance de deux entités (par exemple, Newcombe *et al.* 1959; Fellegi et Sunter 1969; Newcombe 1988; Copas et Hilton 1990). On a moins parlé d'autres aspects du couplage d'enregistrements, comme la manière de traiter les cas de concordance imparfaite de zones d'information, et des conséquences de l'utilisation combinée de diverses approches (traitements).

¹ Thomas R. Belin, Département of Biomathematics, UCLA School of Medicine, Los Angeles, CA, 90024-1766, U.S.A.

Dans certaines circonstances, un identificateur personnel comme le numéro de sécurité sociale pourra servir de critère pour le couplage. Or, on ne connaît pas toujours ces identificateurs, et même si on les connaissait, il faudrait probablement s'appuyer sur d'autres données repères pour un grand sous-ensemble de cas (voir, par exemple, Rogot, Sorlie et Johnson 1986). Cet article décrit une vaste expérience factorielle dans laquelle nous avons comparé diverses méthodes permettant d'apparier des enregistrements du recensement à des enregistrements d'une enquête postcensitaire (EP). Comme le numéro de sécurité sociale n'est pas demandé dans un recensement, nous sommes devant une situation où l'évaluation du degré de concordance repose sur plusieurs variables. Nous cherchons à répondre à deux questions en particulier:

- (1) Quels sont les principaux facteurs qui influent sur la précision du couplage d'enregistrements?
- (2) Quelle combinaison de facteurs donne les meilleurs résultats dans la pratique?

Au delà du fait que cette étude tente de répondre à ces questions par une comparaison du recensement et de l'enquête postcensitaire, son plus grand mérite est probablement de faire valoir que l'analyse des méthodes de couplage d'enregistrements doit se faire au moyen d'expériences menées soigneusement. Si de nombreux facteurs sont laissés sous le contrôle de l'opérateur du programme, il y a peu de chances que l'on comprenne toute la complexité d'un algorithme d'appariement en faisant varier un seul facteur à la fois (ou pire encore, en ne faisant aucune

- NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: methods for health and statistical studies, administration and business*. Oxford: Oxford Medical Publications.
- NEWCOMBE, H.B., FAIR, M.E., et LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1204.
- SCHUBERT, F., et WINKLER, W.E. (1991). An error model for regression analysis of data files that are computer matched. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.
- SCHNATTER, A.R., ACQUIVELLA, J.F., THOMPSON, F.S., DONALESKI, D., et THERIAULT, G. (1990). An analysis of death ascertainment and follow-up through Statistics Canada's Mortality Data Base system. *Canadian Journal of Public Health*, 81, 60-65.
- SHANNON, H.S., JULIAN, J.A., et ROBERTS, R.S. (1984). A mortality study of 11,500 nickel workers. *Journal of the National Cancer Institute*, 73, 1251-1258.
- SMITH, M.E., et NEWCOMBE, H.B. (1982). Use of the Canadian Mortality Data Base for epidemiological follow-up. *Canadian Journal of Public Health*, 73, 39-46.
- SMITH, M.E., et SILINS, J. (1981). Generalized Iterative Record Linkage System. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.
- STATISTIQUE CANADA, (1991a). Canadian Farm Operators' mortality study general work plan. Rapport interne de la section de la recherche sur l'hygiène du travail et de l'environnement. Statistique Canada, Ottawa.
- STATISTIQUE CANADA, (1991b). Données non-publiées.
- TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.
- WIGLE, D.T., SEMENCIW, R.M., WILKINS, K., RIEDEL, D., RITTER, L., MORRISON, H.I., et MAO, Y. (1990). Mortality study of Canadian male farm operators: Non-Hodgkin's lymphoma mortality and agriculture practices in Saskatchewan. *Journal of the National Cancer Institute*, 82, 575-581.
- WINKLER, W.E. (1988). Using the E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.
- WINKLER, W.E., et THIBAUDEAU, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division, rapport technique.

REMERCIEMENTS

Nous tenons à remercier Martha Fair, de Statistique Canada, et le docteur Howard Morrison, de Santé et Bien-être social Canada, pour les précieux commentaires que ces personnes ont exprimés sur cet article. Nous remercions également les deux réviseurs anonymes pour leurs nombreux conseils pratiques.

BIBLIOGRAPHIE

ASHMORE, J.P., et GROGAN, D. (1985). The National Dose Registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.

ASHMORE, J.P., KREWSKI, D., et ZIELINSKI, J.M. (1993). National Dose Registry Study. *European Journal of Cancer*, soumis.

BELIN, T.R. (1989). Results from evaluation of computer matching. Note de service, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R., et RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

BOX, G., et COX, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26, 211-246.

CARPENTER, M., et FAIR, M.E. (Éds.) (1990). *Canadian Epidemiology Research Conference - 1989: Proceedings of Record Linkage Sessions and Workshop*. Ottawa, Ontario: Ottawa Select Printing.

COPAS, J.B., et HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society*, Series A, 153, 287-320.

DEMPSTER, A.D., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via EM algorithm, (avec discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

FAIR, M.E. (1989). Studies and references relating to uses of the Canadian Mortality Data Base. Rapport de la section de la recherche sur l'hygiène du travail et de l'environnement, division de santé, Statistique Canada, Ottawa.

FAIR, M.E., et LALONDE, P. (1988). Identificateurs manquants et justesse de l'observation suivie. *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada, Ottawa, 111-125.

FELLECI, I.P., et SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HILL, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy. Release 2.7. Rapport de la sous-division de la recherche et des systèmes généraux, division des services et développement informatiques, Statistique Canada, Ottawa.

HOWE, G.R., et LINDSAY, J. (1981). A Generalized Iterative Record Linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.

HOWE, G.R., et LINDSAY, J. (1983). A follow-up study of a ten-percent sample of the Canadian Labor Force. I. Cancer mortality in males, 1965-73. *Journal of the National Cancer Institute*, 70, 37-44.

HOWE, G.R., FRASER, D., LINDSAY, J., PRESNALL, B., et YU, S.Z. (1983). Cancer mortality (1965-77) in relation to diesel fume and coal exposure in a cohort of retired railway workers. *Journal of the National Cancer Institute*, 70, 1015-1019.

HOWE, G.R., NAIR, R.C., NEWCOMBE, H.B., MILLER, A.B., BURCH, J.D., et ABBATT, J.D. (1987). Lung cancer mortality (1950-80) in relation to radon daughter exposure in a cohort of workers at the Eldorado Port radium uranium mine: Possible modification of risk by exposure rate. *Journal of the National Cancer Institute*, 79, 1255-1260.

HOWE, G.R., et SPASOFF, R.A. (Éds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. Toronto: University of Toronto Press.

JARO, M.A., (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JABINE, T.B., et SCHUBERT, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.

JORDAN-SIMPSON, D.A., FAIR, M.E., et POLIQUIN, C. (1990). Etude des exploitants agricoles canadiens: Méthodologie. *Rapports sur la santé*, N° 83-003 au catalogue, Statistique Canada, 2, 141-155.

KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. Rapport présenté à la réunion de l'American Statistical Association, août 1986.

LABOSSIERE, G. (1986). Confidentiality and access to data: the practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press.

MAO, Y., SEMENCIW, R., MORRISON, H., KOCH, M., HILL, G., FAIR, M., et WIGLE, D. (1988). Survival rates among patients with cancer in Alberta in 1974-78. *Canadian Medical Association Journal*, 138, 1107-1113.

MENG, L., et RUBIN, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899-911.

MORRISON, H.I., SEMENCIW, R.W., MAO, Y., et WIGLE, D.T. (1988). Cancer mortality among a group of fluorospar miners exposed to radon progeny. *American Journal of Epidemiology*, 128, 1266-1275.

NETER, J., MAYNES, E.S., et RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., et ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medicine*, 13, 157-169.

4. QUESTIONS TOUCHANT L'ANALYSE DES ENSEMBLES DE DONNÉES COUPLÉES

Peu d'études ont été faites pour déterminer l'effet du couplage d'enregistrements sur les résultats de l'analyse de régression. Neter *et al.* (1965) ont reconnu que les erreurs introduites durant l'appariement pouvaient nuire aux analyses qui se font à partir des fichiers qui viennent d'être couplés. Supposons que les valeurs vraies d'une variable aléatoire étudiée sont inscrites dans un fichier constitué de N enregistrements. Désignons par Y_i la valeur vraie pour l'enregistrement $i = 1, \dots, N$. Ce fichier est couplé à un autre fichier, qui contient des renseignements personnels et selon lequel une valeur Z_i est attribuée à l'enregistrement $i = 1, \dots, N$. En supposant que toutes les erreurs d'appariement ont la même probabilité, nous avons

$$Z_i = \begin{cases} Y_i & \text{avec une probabilité } p \\ Y_j & \text{avec une probabilité } q \text{ (} j \neq i \text{),} \end{cases}$$

$$\text{où } p + (N - 1)q = 1.$$

Neter *et al.* (1965) se sont servis de ce modèle pour étudier l'incidence des erreurs d'appariement sur la moyenne et la variance empiriques de la variable Z . Ils ont aussi étudié l'effet des erreurs d'appariement sur la corrélation entre Z et une autre variable aléatoire, X , incluse dans le même fichier et sur les estimations des paramètres de la régression de Z_i par rapport à X_i . Leurs conclusions étaient les suivantes: 1) l'estimation de la moyenne de Z est non biaisée pour la moyenne des X_i ; 2) si W est corrélié positivement avec X , la variance résiduelle d'une régression de Z par rapport à X sera plus grande que la variance d'une régression de X par rapport à W ; et 3) la pente de la droite de régression sera sous-estimée si Z est utilisée au lieu de X .

Belin et Rubin (1991) et Winkler et Thibaudreau (1991) étudient des cadres théoriques, des algorithmes de calcul et des logiciels pour estimer les probabilités d'appariement. Ces recherches ont incité Scheuren et Winkler (1991) à mettre à jour les travaux de Neter *et al.* (1965). Ils ont utilisé le modèle

$$Z_i = \begin{cases} Y_i & \text{avec probabilité } p_i \\ Y_j & \text{avec probabilité } q_{ij} \text{ (} j \neq i \text{),} \end{cases}$$

où $p_i + \sum_{j \neq i} q_{ij} = 1$, pour étudier l'incidence des erreurs d'appariement sur les estimations des coefficients β dans le modèle de régression linéaire

$$Y = X\beta + \epsilon.$$

L'effet des erreurs d'appariement sur le modèle de régression ci-dessus peut être exprimé par l'équation

5. ANALYSE

Le couplage d'enregistrements est une méthode intéressante pour examiner les liens entre l'exposition professionnelle et l'état de santé au moyen de bases de données existantes. Cependant, des erreurs de couplage peuvent se produire pour diverses raisons: erreurs de codage, changement dans les identificateurs, données manquantes, pouvoir de différenciation insuffisant des identificateurs. Les taux d'erreur dépendent de la quantité de renseignements personnels, comme le montre l'étude des exploitants agricoles. Dans cette étude, le taux d'erreur était moins élevé avec le couplage du questionnaire de recensement complet, qui contenait plus de renseignements personnels que le questionnaire abrégé. Il est donc important de disposer de renseignements personnels de qualité pour effectuer un couplage d'enregistrements.

On a porté assez peu d'attention à l'incidence des erreurs de couplage sur les inférences statistiques basées sur des études par couplage d'enregistrements. Ce genre d'erreur peut introduire un biais dans les estimations des mesures d'association des variables de santé et des variables environnementales, comme les coefficients de régressions. Des recherches sont en cours dans le but d'analyser l'effet de ces erreurs sur les résultats des études épidémiologiques présentées dans cet article.

On a porté assez peu d'attention à l'incidence des erreurs de couplage sur les inférences statistiques basées sur des études par couplage d'enregistrements.

Au lieu d'utiliser la paire formée de la variable indépendante et de la variable dépendante, (X_i, Y_i) , pour ajuster le modèle, on se sert de la paire formée de la variable indépendante et de la variable dépendante couplée, (X_i, Z_i) . Notons que la variable dépendante couplée peut être exprimée sous la forme $Z = Y + B$, les coefficients étant estimés par l'expression

$$B_i = (p_i - 1)Y_i + \sum_{j \neq i} q_{ij}Y_j.$$

où le terme de biais est défini

$$E(Z_i) = Y_i + B_i,$$

3.2.2 Méthode de couplage d'enregistrements

Les variables d'identification ont changé plusieurs fois depuis la création du FDN, ce qui complique par moments la reconstitution du passé dosimétrique des individus. À cause de cela et d'autres difficultés du même genre, on se sert du numéro d'assurance sociale comme variable d'identification clé depuis 1977.

Plusieurs couplages ont été nécessaires pour rassembler les identificateurs personnels, les antécédents dosimétriques et les données sur les décès appropriés.

a) Couplage à l'intérieur des fichiers dosimétriques. Depuis 1984, Statistique Canada exécute des fusions dynamiques avec son fichier LDHS afin de regrouper les enregistrements dosimétriques; ces fusions ont pour effet de réduire le nombre d'enregistrements fragmentaires et de réunir les enregistrements dans un dossier dosimétrique complet pour chaque individu étudié. Le fichier dérivé des couplages internes indique quels enregistrements du FDN semblent se rapporter au même individu.

b) Couplage avec la BCDM. La cohorte du FIC, qui avait déjà fait l'objet d'un couplage interne, a été couplée avec les données sur la mortalité (couplage à deux fichiers). En couplant ces deux fichiers, il est possible d'évaluer le risque ultérieur de décès pour les membres de la cohorte. Dans cette étude, la BCDM permettrait de connaître la cause du décès, l'année et le lieu du décès, de même que le lieu et l'année de naissance.

c) Fichier d'analyse. On a apparié des données de trois fichiers – FIC, BCDM et LDHS – pour créer un enregistrement complet pour chaque membre de la cohorte étudiée. Chaque enregistrement indique, si possible, le mois et l'année de naissance, le sexe, les données sur les décès énumérées ci-dessus, le poids de couplage avec les données sur le décès, et les antécédents dosimétriques. Les enregistrements du FIC ou du fichier dosimétrique qui ne pouvaient être appariés étaient soumis à un examen minutieux.

3.2.3 Choix des seuils

En ce qui a trait au couplage du fichier de cohorte avec la BCDM, le choix du seuil s'est effectué de la même manière que dans l'étude des exploitants agricoles canadiens. Premièrement, on a déterminé les poids des cas indéterminés. Tous les cas indéterminés qui avaient un poids inférieur à — 30 étaient classés comme des non-liens. On comptait 4,429 femmes et 8,686 membres de la cohorte avec un poids de couplage supérieur à cette valeur. On a tiré un échantillon dans ce groupe de personnes et on l'a traité manuellement en examinant les certificats de décès pour déterminer si les liens observés étaient des liens authentiques ou des non-liens. Le seuil a été choisi de manière à correspondre au poids de couplage pour lequel le nombre de faux liens positifs est égal au nombre de faux liens négatifs pour les femmes et pour les hommes, pris séparément (figure non insérer).

Pour illustrer l'effet de l'addition de renseignements, on peut construire un tableau semblable pour les questions nées complètes (tableau 1). Les taux d'erreur pour les faux positifs et les faux négatifs sont 4,3% ((13/299) × 100) et 0,1% ((15/18,511) × 100) respectivement, pour un taux global de 4,4%. On peut donc réduire les taux d'erreur en augmentant la quantité de renseignements personnels.

3.2 Étude de mortalité fondée sur le Fichier dosimétrique national

Le Fichier dosimétrique national (FDN) du Canada contient des données sur l'exposition professionnelle au rayonnement pour environ 255,000 Canadiens; ces données remontent aussi loin que 1951. Le FDN a été couplé récemment à la BCDM. L'étude de mortalité fondée sur le Fichier dosimétrique national a pour but d'établir un lien entre la surmortalité attribuable au cancer et d'autres affections et l'exposition professionnelle à de faibles niveaux de rayonnement ionisant (Ashmore *et al.* 1993).

3.2.1 Définition de la cohorte

La cohorte se compose de tous les travailleurs qui ont fait l'objet d'une surveillance pour rayonnement ionisant (y compris les produits de filiation du radon et du tritium) et qui avaient un dossier dans le Fichier dosimétrique national au 31 décembre 1983. Le FDN contient les fiches de presque tous les travailleurs exposés à des rayonnements ionisants au Canada qui font l'objet d'une surveillance; certaines fiches contiennent des données qui remontent à 37 ans. Le FDN comprend en outre 80 catégories d'emploi, qui vont du travailleur de centrale nucléaire au dentiste en passant par le radiologiste d'hôpital. En tout, 248,940 personnes formaient la cohorte à l'étude.

Selon le genre de rayonnement et le degré d'exposition prévu pour des catégories d'emploi particulières, des données sont recueillies annuellement, trimestriellement, mensuellement ou à la quinzaine. À chaque année, un indice global de la dose absorbée annuellement par chaque individu est inscrit dans le Lifetime Dose History System (LDHS). C'est à partir des données contenues dans le LDHS que l'on pourra tenter d'établir des liens entre l'exposition professionnelle aux rayonnements et l'état de santé.

Les données individuelles contenues dans le LDHS permettent aussi de calculer la dose reçue cumulée pour chaque personne. Bien que le degré d'exposition des personnes ne sera pas le même à chaque année, on peut calculer une dose annuelle moyenne pour les individus en divisant la dose reçue cumulée par le nombre d'années qui se sont écoulées depuis le moment de la première exposition. Les analyses statistiques peuvent se faire en fonction de la dose reçue cumulée, de la dose annuelle moyenne ou des doses annuelles inscrites dans le LDHS. Jusqu'en 1986, les renseignements personnels tels que le nom de famille, le prénom, le sexe, l'année de naissance et les numéros d'identification attribués aux fichiers des individus étaient stockés dans un fichier indépendant appelé fichier d'identification central (FIC) (Ashmore et Grogan 1985).

c) Fichier d'analyse. La base de données de la cohorte des exploitants agricoles a été couplée avec la BCDM au moyen du CBI. Le fichier ainsi obtenu contenait des données socio-démographiques de même que des données sur l'exposition aux risques environnementaux et sur la mortalité, et pouvait donc se prêter à l'analyse.

3.1.3 Choix des seuils

Il fallait déterminer des valeurs de seuil pour chacun des trois couplages effectués en vue de constituer le fichier d'analyse et pour les couplages basés sur le questionnaire abrégé et le questionnaire complet. En ce qui a trait au couplage avec des bases de données sur la mortalité, Statistique Canada a fixé les seuils à l'aide d'une méthode basée sur un échantillon. Cette méthode est illustrée pour le couplage de la base de données de la cohorte des exploitants agricoles avec la base de données sur la mortalité (en ce qui concerne ceux qui ont rempli le questionnaire de recensement abrégé).

On a prélevé un échantillon d'environ 10% parmi les questionnaires abrégés qui avaient été remplis par la cohorte des exploitants agricoles (Statistique Canada 1991a). On a ensuite déterminé les liens de deux manières: par le système de CBI de Statistique Canada et par une technique manuelle qui utilise les données de registres des décès. Après quoi on a comparé les résultats des couplages en supposant que les couplages réalisés par la technique manuelle produisaient des concordances vraies. La figure 2 représente le nombre de faux positifs et de faux négatifs qui correspondent à une série de seuils comme

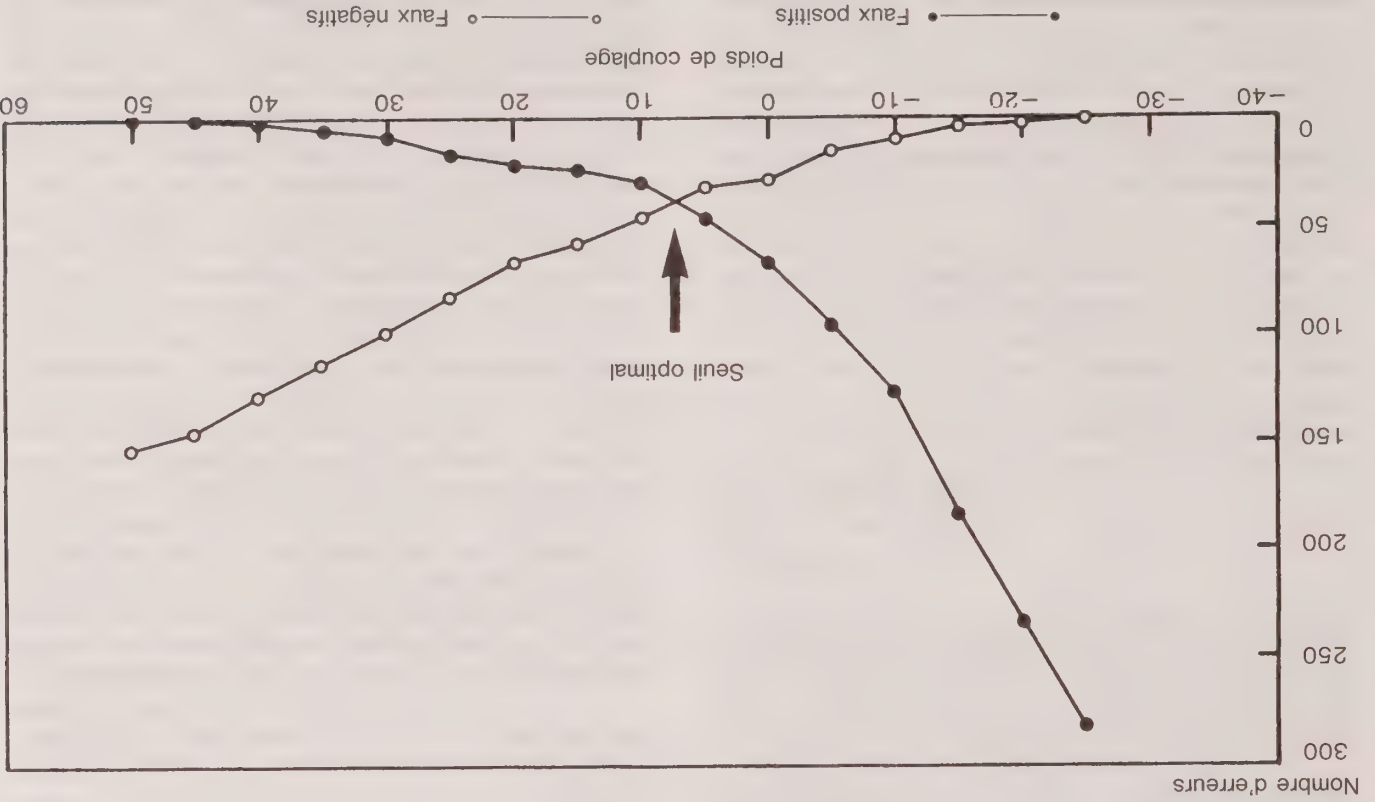


Tableau 1
Comparaison d'enregistrements couplés et non couplés au moyen du CBI et de la technique manuelle; recensement pour l'étude des exploitants agricoles

Couplage	Technique manuelle		Couplage d'enregistrements informatisés (CEI)	Questionnaire abrégé		Couplages	Non couplés	Total
	Non couplés	Couplés		Couplés	Questionnaire complet			
Couplages	417	36	453	417	38	455	453	453
Non couplés	20,809	20,845	20,847	20,809	20,845	20,845	20,847	21,300
Couplages	286	13	299	286	13	286	299	299
Non couplés	18,498	18,513	18,513	18,498	18,513	18,511	18,513	18,812
Total	301	301	301	301	301	301	301	301

poids de couplage. Le seuil a été choisi de manière à minimiser le nombre total de faux positifs et de faux négatifs. La valeur-seuil en question est w_0 , point d'intersection des deux courbes. Le taux d'erreur pour les faux positifs est estimé à $(36/453) \times 100 = 7.9\%$, tandis que celui pour les faux négatifs est estimé à $(13/20,847) \times 100 = 0.2\%$, pour un taux d'erreur global de 8.1% (tableau 1).

Figure 2. Faux liens positifs et faux liens négatifs, étude de mortalité des exploitants agricoles canadiens: questionnaire abrégé

de la cohorte des exploitants agricoles à la BCDM. Mais avant d'effectuer ce troisième couplage, il a fallu consulter le fichier de la cohorte des exploitants agricoles. Les données socio-démographiques ont été tirées du recensement de la population de 1971 tandis que les données sur les pratiques agricoles provenaient du recensement de l'agriculture de 1971. La base de données du recensement de la population contient des enregistrements pour chaque personne au Canada et elle est constituée de données recueillies au moyen de deux types de questionnaires: un questionnaire abrégé et un questionnaire complet. Ce dernier exigeait plus de renseignements que le premier et était distribué aléatoirement au tiers des ménages. Le questionnaire du recensement de l'agriculture était envoyé à toutes les exploitations agricoles.

Le nom des exploitants agricoles n'apparaît pas comme tel dans le fichier du recensement de l'agriculture ni dans celui du recensement de la population. Leurs nom et adresse figurent dans le registre central des fermes, créé pour servir de liste d'envoi pour les questionnaires sur l'agriculture. La BCDM fait état de tous les décès enregistrés par les provinces et les territoires depuis 1950 et est stockée sous une forme standard dans un ordinateur à Statistique Canada (Smith et Newcombe 1982). Entre 1950 et 1987, 5.9 millions de décès ont été enregistrés dans la BCDM. La base contient des renseignements personnels, plus la date et l'endroit du décès de même que la cause du décès, codée selon la Classification internationale des maladies (CIM).

La Loi sur la statistique garantit la confidentialité de tous les enregistrements contenus dans la BCDM et les bases de données des recensements de la population et de l'agriculture. Comme nous l'avons mentionné plus tôt, toutes les études qui nécessitent des couplages avec ces bases de données doivent être soumises à un processus d'examen et d'approbation minutieux avant d'être réalisées, et les fichiers couplés qui contiennent des renseignements personnels demeurent sous la garde de Statistique Canada.

a) **Suivi.** Le registre central des fermes de 1971 a été couplé à celui de 1981 au moyen du CBI pour déterminer si les agriculteurs qui figuraient sur la liste en 1971 étaient encore vivants en 1981. Cette information était ajoutée dans le registre central des fermes de 1971 afin d'accroître la probabilité d'un couplage fructueux avec la BCDM.

b) **Base de données de la cohorte des exploitants agricoles.** Le registre central des fermes de 1971, enrichi des données de l'étape précédente, a été fusionné au fichier du recensement de l'agriculture afin d'inclure des noms dans la base de données de la cohorte. Cette opération était nécessaire pour effectuer le couplage avec la base de données sur la mortalité. Le fichier ainsi obtenu était ensuite couplé au fichier du recensement de la population à l'aide du CBI pour constituer la base de données de la cohorte des exploitants agricoles.

Après avoir estimé le coefficient de mixité, λ , à l'aide de résultats d'opérations d'appariement antérieures, on peut ajuster le modèle ci-dessus au moyen de poids issus de la procédure de couplage. On peut ensuite se servir du modèle ajusté pour estimer le taux d'erreur de l'algorithme de couplage d'enregistrements, étant donné un seul participant. L'erreur type correspondante est aussi estimée au moyen de l'algorithme SEM, par lequel est estimée la covariance des estimations de paramètres produites par l'algorithme EM (Meng et Rubin 1991).

3. EXEMPLES DE GRANDES ÉTUDES PAR COUPLAGE D'ENREGISTREMENTS

3.1 Étude des exploitants agricoles canadiens

L'étude des exploitants agricoles canadiens a été créée dans le but d'analyser les liens possibles entre des causes de décès chez les exploitants agricoles et diverses variables socio-démographiques et agricoles, en particulier le rapport entre l'emploi de pesticides et la mortalité. Les données sur la mortalité provenaient de la BCDM, tandis que celles sur les variables socio-démographiques et agricoles étaient tirées du recensement de la population et du recensement de l'agriculture. Comme les bases de données des recensements ne contenaient pas d'informations sur l'exposition aux pesticides proprement dite, on s'est servi de variables substitutives comme le nombre d'acres traités aux insecticides ou aux herbicides et le coût des produits chimiques agricoles. Le fichier d'analyse contenant les données pertinentes a été construit à l'aide d'une méthode de couplage probabiliste.

3.1.1 Définition de la cohorte

La cohorte se compose de tous les agriculteurs de sexe masculin qui répondaient à la définition d'exploitant agricole dans le recensement de 1971. L'exploitant agricole est la personne responsable des décisions quotidiennes prises pour la bonne marche de l'exploitation agricole. Il n'est pas nécessairement propriétaire de l'exploitation; il peut en être locataire ou agir à titre de gérant engagé. Un seul exploitant est désigné pour chaque exploitation agricole. Selon la définition que l'on en donnait dans le recensement de 1971, une ferme est une exploitation agricole qui a une superficie d'un acre ou plus et dont les ventes de produits agricoles sont de \$50 ou plus. La cohorte des exploitants agricoles comptait 326,000 individus de sexe masculin (Jordan-Simpson *et al.* 1990). On en a fait le suivi (au point de vue de la mortalité) jusqu'en 1987.

3.1.2 Méthode de couplage d'enregistrements

Le fichier d'analyse de l'étude des exploitants agricoles canadiens a été constitué par suite de trois couplages indépendants, le dernier d'entre eux étant le couplage du fichier

Dans beaucoup d'applications toutefois, la solution manuelle n'est pas pratique, surtout lorsqu'il y a un grand nombre de cas indéterminés. Dans ces circonstances, on peut déterminer un seuil unique, $w_i = w_i^n$, de sorte que deux possibilités seulement existent: les paires dont le poids est plus grand que w_i sont reconnues comme des liens; celles dont le poids est plus petit que w_i sont reconnues comme des cas indéterminés. On peut s'appuyer sur des données supplémentaires pour prendre une décision sur les cas indéterminés. Dans beaucoup d'applications toutefois, la solution manuelle n'est pas pratique, surtout lorsqu'il y a un grand nombre de cas indéterminés. Dans ces circonstances, on peut déterminer un seuil unique, $w_i = w_i^n$, de sorte que deux possibilités seulement existent: les paires dont le poids est plus grand que w_i sont reconnues comme des liens; celles dont le poids est plus petit que w_i sont reconnues comme des cas indéterminés.

Le choix de la limite w_i n'est pas sans difficultés. Les méthodes existantes reposent sur une connaissance des taux d'erreur de couplage, qui sont estimés soit par la résolution manuelle d'un échantillon (ou d'une population) de cas indéterminés, soit par l'application d'une méthode analytique. La première possibilité représente une méthode fondée sur un échantillon puisqu'elle implique une collecte de données pour estimer le taux d'erreur de couplage.

Les taux d'erreur de couplage dépendent des seuils fixés. Plus l'écart est grand entre les limites inférieure et supérieure, plus le nombre de cas indéterminés est élevé. Si nous avons un seuil unique, le nombre de "faux négatifs" augmente et le nombre de "faux positifs" diminue à mesure que le seuil s'accroît.

L'emploi d'une méthode simple fondée sur un échantillon pour le choix du seuil nécessite l'application d'une phase pilote. Premièrement, on tire un échantillon d'enregistrements dans le plus petit des deux fichiers qui doivent faire l'objet d'un couplage. Deuxièmement, on établit des concordances par une technique manuelle et au moyen d'un système de couplage probabiliste informatisé. Troisièmement, en supposant que les concordances établies manuellement sont des concordances vraies, le seuil choisi correspond au poids pour lequel la somme des faux positifs et des faux négatifs. Malgré tout, des erreurs de couplage pourraient encore se produire dans les opérations manuelles à cause d'erreurs de codage, du pouvoir de différenciation insuffisant des identificateurs utilisés ou d'autres problèmes de couplage.

Pour estimer les taux d'erreur d'un système de CEI, on peut construire un tableau de contingence comme celui-ci :

On peut alors estimer le taux de faux positifs (FP) et

CEI	Couplage	n_{11}	n_{12}
	Non-couplage	n_{21}	n_{22}

CEI	Non-couplage	Couplage	Non-couplage
	Système manuel		
		Couplage	Non-couplage
		n_{11}	n_{12}
		n_{21}	n_{22}

$$\frac{n_{11} + n_{12}}{n_{12}} = \mathbf{FP}$$

Felleget et Sunter (1969) soulignent que les taux d'erreur rattachés à des seuils donnés sont une fonction des probabilités de concordance pour les liens et les non-liens. Les estimations des probabilités de concordance peuvent donc servir à déterminer les seuils. Cette approche est aussi examinée par Jaro (1989).

En ce qui concerne les systèmes de couplage d'entre-

seuls ressemblent, à quelques différences près, à celles appliquées dans la méthode à échantillon décrite plus haut. L'essentiel de la méthode basée sur un modèle consiste à ajuster des modèles en vue d'estimer les probabilités conditionnelles définies en (1) et (2) et d'estimer le taux d'erreur au moyen du logit de deux probabilités conditionnelles estimées. Cette méthode prévoit l'utilisation de l'algorithme EM pour estimer les probabilités conditionnelles m_i et u_i définies en (1) et (2) pour la zone i de l'enregistrement en supposant l'indépendance des comparaisons entre les zones,

$$\left\{ \begin{array}{l} 0 \text{ s'il y a non-concordance des zones } i \text{ pour la} \\ \text{paire d'enregistrements } j. \\ 1 \text{ s'il y a concordance des zones } i \text{ pour la} \\ \text{paire d'enregistrements } j \end{array} \right\} = \gamma_i^j$$

$$Pr(\gamma^i|U) = \Pi_n^{i=1} n_i^{1-n_i} (1 - n_i)^{1-n_i},$$

$$\left. \begin{array}{l} 0 \text{ s'il y a non-concordance des zones } i \text{ pour la} \\ \text{paire d'enregistrements } j. \\ 1 \text{ s'il y a concordance des zones } i \text{ pour la} \\ \text{paire d'enregistrements } j \end{array} \right\} = \gamma_i^j$$

2.3 Sources d'erreur

Dans le couplage d'enregistrements, il existe un certain nombre de sources d'erreur qui peuvent causer un mauvais appariement d'enregistrements. Des erreurs de codage (par ex. date de naissance erronée) peuvent se produire lorsque les enregistrements sont entrés dans les bases de données. Il peut aussi y avoir des différences de code, par exemple des versions différentes du prénom ou du nom de famille. Outre les erreurs de codage et les différences de code, les données manquantes, spécialement celles qui ont trait à des identificateurs importants, feront s'accroître sensiblement le taux d'erreur pour le couplage d'enregistrements (Fair et Lalonde 1981). L'existence d'enregistrements en double, c'est-à-dire qu'un enregistrement d'un fichier est couplé à plus d'un enregistrement de l'autre fichier, peut aussi causer des erreurs de couplage (Jabine et Scheuren 1986). C'est pourquoi les systèmes de CBI doivent inclure des règles qui permettent les appariements multiples.

Une façon d'accroître la fiabilité de l'identificateur du nom de famille est d'utiliser un système de codage phonétique. Par exemple, deux observations d'un identificateur, ANDERSON et ANDERSEN, seront recodées ANDAR au moyen du New York State Intelligence and Identification System (NYSIS) (Newcombe 1988). Cela aura pour conséquence d'atténuer l'effet des variations de nom sur le couplage des enregistrements. Cependant, en comprimant le nom, on risque de réduire le pouvoir de différenciation puisque deux noms différents peuvent avoir le même code NYNIS. Du même coup, la probabilité d'effectuer des couplages incorrects s'accroît (Newcombe 1988).

Le prénom peut se présenter sous diverses formes, selon les bases de données. Par exemple, Joseph et Joe, Cynthia et Cindy, David et Dave. Newcombe *et al.* (1992) étudient des façons de mettre à profit les connaissances acquises sur les variations de prénoms afin d'accroître la probabilité d'effectuer des couplages exacts.

Lorsqu'il s'agit de grands fichiers, il n'est plus commode de comparer tous les couples d'enregistrements possibles. Afin de réduire le nombre de comparaisons, on peut répartir les enregistrements des deux fichiers à coupler dans des groupes exhaustifs s'excluant mutuellement et faire les comparaisons à l'intérieur des groupes. Le groupage s'effectue généralement par un tri appliqué aux deux fichiers à l'aide d'une ou de plusieurs variables d'identification. L'inconvénient de cette méthode est que les éléments d'une paire qui sont classés dans deux groupes différents ne seront pas comparés l'un à l'autre et seront par conséquent assimilés à des non-concordances. Les paires d'éléments à rapprocher proviendront uniquement des enregistrements pour lesquels il y concordance des variables de tri. Par conséquent, le nombre de faux liens négatifs augmentera (Newcombe 1987; Jaro 1989). Une

2.4 Choix du seuil et estimation du taux d'erreur

Après que des poids ont été attribués à toutes les concordances probables, une décision est prise concernant la probabilité que la concordance soit une concordance désignée (c.-à-d. un lien). Suivant la méthode de Fellegi-Sunter, on compare chaque poids à des limites supérieure et inférieure et on prend une décision selon la règle suivante:

$$\text{concordance} = \begin{cases} \text{un lien} & \text{si } w \geq w_u \\ \text{un cas indéterminé} & \text{si } w_l < w < w_u \\ \text{un non-lien} & \text{si } w \leq w_l \end{cases}$$

Dans la règle ci-dessus, w_l et w_u sont les limites inférieure et supérieure pour les poids de couplage; idéalement, elles sont choisies de manière à réduire au maximum le nombre de cas indéterminés, ce qui a pour effet de maintenir, voire d'abaisser le taux d'erreur de classification pour les deux types d'erreur (lien authentique classé comme non-lien, et non-lien authentique classé comme lien).

bonne variable de groupage est une variable qui repose sur des groupes qui comptent à peu près le même nombre d'enregistrements (Jaro 1989).

Dans la plupart des applications de la méthode de Fellegi-Sunter, les résultats des comparaisons faites pour différentes zones d'appariement sont supposés indépendants. Kelley (1986) a fait des études de simulation pour analyser la robustesse du système de couplage du Bureau of the Census des E.-U. à l'égard du non-respect de l'hypothèse d'indépendance. Pour certaines populations et certaines variables de couplage, on a constaté que le non-respect de l'hypothèse d'indépendance peut avoir une incidence notable sur les taux d'erreur de couplage.

Newcombe *et al.* (1983) ont comparé la précision de l'appariement informatisé à celle de la recherche manuelle dans une étude de suivi épidémiologique. Ils ont constaté que l'appariement informatisé était plus avantageux que la recherche manuelle et moins susceptible de favoriser un couplage (nécessairement faux) avec des enregistrements qui n'ont aucun rapport avec la population à l'étude. Dans les deux cas, la précision dépendait largement de la quantité de renseignements personnels contenus dans les enregistrements qui faisaient l'objet du couplage. Fair et Lalonde (1987) arrivent aux mêmes conclusions après avoir étudié l'effet de la présence (ou de l'absence) de divers identificateurs sur les taux d'erreur de couplage.

Schnatter *et al.* (1990) ont vérifié si le système de CBI utilisé à Statistique Canada permet d'évaluer avec précision un nombre de décès. On a comparé le nombre de décès survenus parmi une cohorte de 17,446 travailleurs de l'industrie du raffinage du pétrole avec le nombre de décès établi au moyen d'un couplage à la BCDM. Le système de CBI de Statistique Canada avait signalé 98% des décès survenus au Canada.

L'enregistrement 1 du fichier A concorde avec l'enregistrement 1 du fichier B et l'enregistrement 2 de la base de données A concorde avec l'enregistrement 3 du fichier B. L'enregistrement 3 du fichier A ne correspond à aucun des enregistrements du fichier B et l'enregistrement 2 du fichier B ne correspond non plus à aucun des enregistrements du fichier A.

Si les enregistrements renferment des identificateurs uniques qui ont été attribués avec précision, l'appariement sera une opération élémentaire. Le numéro d'assurance sociale est un exemple d'identificateur propre à un individu. Cependant, il n'existe pas toujours d'identificateurs uniques; dans ces circonstances, on ne peut effectuer de quelconque de couplage probabiliste (voir section 2.3). Grâce à ce mode de couplage, on peut calculer la probabilité d'une concordance et utiliser un système de poids de couplage pour déterminer les liens et les non-liens.

2.2 Système de couplage d'enregistrements informatisé (CEI)

Dans un mode de couplage probabiliste, la décision préliminaire d'apparier ou non dépend d'un poids établi par suite de la comparaison de composantes de deux enregistrements (Newcombe 1988). Ce poids reflète la probabilité que cette paire d'enregistrements constitue un vrai lien; plus le poids est élevé, plus la paire est susceptible de constituer un vrai lien. Le poids dépend ordinairement de la probabilité d'appariement calculée au moment de la comparaison de deux enregistrements.

Dans l'expression ci-dessus, M désigne l'événement – c'est-à-dire la concordance de deux enregistrements – et $\{A, B, \dots, Z\}$ sont les résultats de la comparaison d'identificateurs personnels. Le poids w est défini par le

$$w = \log_2 \left\{ \frac{P(M|AB, \dots, Z)}{P(M|AB, \dots, Z)} \right\}$$

$$= w_a + w_b + \dots + w_z + w,$$

où

$$w_j = \log_2 \left\{ \frac{P(j|M)}{P(j|M)} \right\}$$

pour tous $j \in \{A, B, \dots, Z\}$, et

$$W = \log_2 \left\{ \frac{P(M)}{P(M)} \right\}.$$

Notons que pour obtenir une valeur absolue, on doit connaître le nombre de concordances vraies et le nombre de non-concordances. Sinon, on ne peut déterminer que la probabilité relative. Le poids calculé par le système de CEI de Statistique Canada est la différence de logits.

On a élaboré des algorithmes visant à attribuer des poids qui reflètent la probabilité de couplage de deux enregistrements; à cette fin, on a posé l'hypothèse que les probabilités de concordance pour chacun des identificateurs pris individuellement sont statistiquement indépendantes (Howe et Lindsay 1981). Toutefois, certains identificateurs peuvent être corrélés, ce qui introduit un biais dans l'attribution du poids global.

Fellegi et Sunter (1969) ont proposé un modèle mathématique dans le but de définir un cadre théorique pour le couplage d'enregistrements. Dans ce modèle, le poids tient compte des probabilités d'erreur pour chaque zone en utilisant un rapport de vraisemblance; le poids w est défini

$$w = \sum_{i \in \{\text{zones}\}} w_i$$

$$w_i = \begin{cases} \log_2 \{m_i/u_i\} & \text{si les zones } i \text{ de deux enregistrements concordent} \\ \log_2 \{(1 - m_i)/(1 - u_i)\} & \text{si les zones } i \text{ de deux enregistrements ne concordent pas,} \end{cases}$$

avec

$$m_i = \Pr\{\text{concordance des zones } i \mid \text{paire d'enregistrements} \in M\} \quad (1)$$

$$u_i = \Pr\{\text{concordance des zones } i \mid \text{paire d'enregistrements} \in U\}. \quad (2)$$

Dans les équations ci-dessus, M est un ensemble de paires d'enregistrements qui sont des concordances et U est un ensemble de paires d'enregistrements qui sont des non-concordances. Les résultats de chaque comparaison de zones sont aussi supposés statistiquement indépendants (Jaro 1989).

Newcombe (1988), Fellegi et Sunter (1969), Tepping (1968) et Copas et Hilton (1990) ont élaboré diverses méthodes – probabilistes ou basées sur un modèle – pour attribuer des poids à des composantes (zones) d'enregistrements. Un système probabiliste comme celui utilisé à Statistique Canada détermine les poids de couplage en calculant le logarithme de la probabilité empirique d'une concordance; d'autres systèmes, basés sur un modèle, utilisent l'algorithme EM (Dempster *et al.* 1977) pour estimer les poids de couplage (Jaro 1989; Belin 1989; Winkler 1988).

Cet article a pour but d'analyser l'utilisation du couplage d'enregistrements informatisés dans des enquêtes épidémiologiques fondées sur des dossiers administratifs touchant la santé et l'environnement. Notre analyse s'intéresse particulièrement à l'effet des faux liens sur les inférences statistiques concernant les risques environnementaux. Dans la section 2, nous étudions des algorithmes pour le couplage d'enregistrements informatisés. La section 3 décrit l'application du couplage d'enregistrements dans des études sur l'exposition professionnelle au rayonnement ionisant et aux produits chimiques agricoles. Dans la section 4, nous penchons sur les questions statistiques que soulève l'analyse de bases de données constituées par le couplage d'enregistrements. Enfin, dans la section 5, nous présentons nos conclusions relativement à l'utilisation du couplage d'enregistrements en épidémiologie du milieu.

2. QUESTIONS RELATIVES AU COUPLAGE D'ENREGISTREMENTS

2.1 Position du problème

Considérons deux fichiers informatiques, **A** et **B**, qui renferment, dans le premier cas, des données sur la santé et, dans le second cas, des données sur l'exposition à des risques environnementaux, pour deux groupes de personnes. Chaque fichier est constitué d'un certain nombre d'enregistrements ou d'"observations" qui renferment chacun un certain nombre de zones ou "composantes". Chaque observation correspond habituellement à un membre de la population. Les zones représentent des attributs, comme le nom, l'adresse, l'âge et le sexe, qui caractérisent les observations. Le couplage d'enregistrements sert à déceler et à coupler les observations de chaque fichier qui corres-

pondent au même individu (figure 1). Dans cet exemple, les algorithmes sophistiqués pour évaluer la probabilité d'une concordance entre deux enregistrements (Hill 1988; Newcombe 1988). Statistique Canada a mis au point un système de CEI appelé CANLINK qui peut effectuer des couplages à l'intérieur d'un même fichier – ou couplages internes – aussi bien que des couplages entre deux fichiers distincts (Howe et Lindsay 1981; Smith et Silins 1981). La confidentialité des enregistrements protégés en vertu de Loi sur la statistique est rigoureusement respectée si ces enregistrements doivent servir dans des études qui nécessitent des couplages. Toutes les études qui nécessitent des couplages avec des bases de données protégées doivent être soumises à un processus d'examen et d'approbation ministériels avant d'être réalisées. Tous les fichiers couplés qui renferment des données personnelles demeurent sous la garde de Statistique Canada (Labossière 1986).

Les études par couplage d'enregistrements offrent plusieurs avantages par rapport aux études épidémiologiques classiques. En utilisant des bases de données administratives existantes, on n'a plus besoin de recueillir de nouvelles données pour des études sur la santé. En ayant accès à des bases de données existantes, on peut souvent obtenir de grands échantillons avec assez de facilité. Suivant la nature des bases de données utilisées, le couplage d'enregistrements est une façon peu coûteuse d'examiner de nombreuses associations possibles dans des études épidémiologiques. Le couplage d'enregistrements a aussi des inconvénients. Des erreurs d'appariement peuvent se produire à cause de différences de codage ou de la non-unicité des identificateurs. On a généralement peu de contrôle sur les données recueillies et dans beaucoup de cas, les opérations de suivi peuvent être infructueuses. En outre, les études par couplage d'enregistrements présentent les mêmes lacunes que les études épidémiologiques classiques, par exemple possibilité de biais, confusion, et difficulté à reconnaître de faibles associations entre le milieu et la santé.

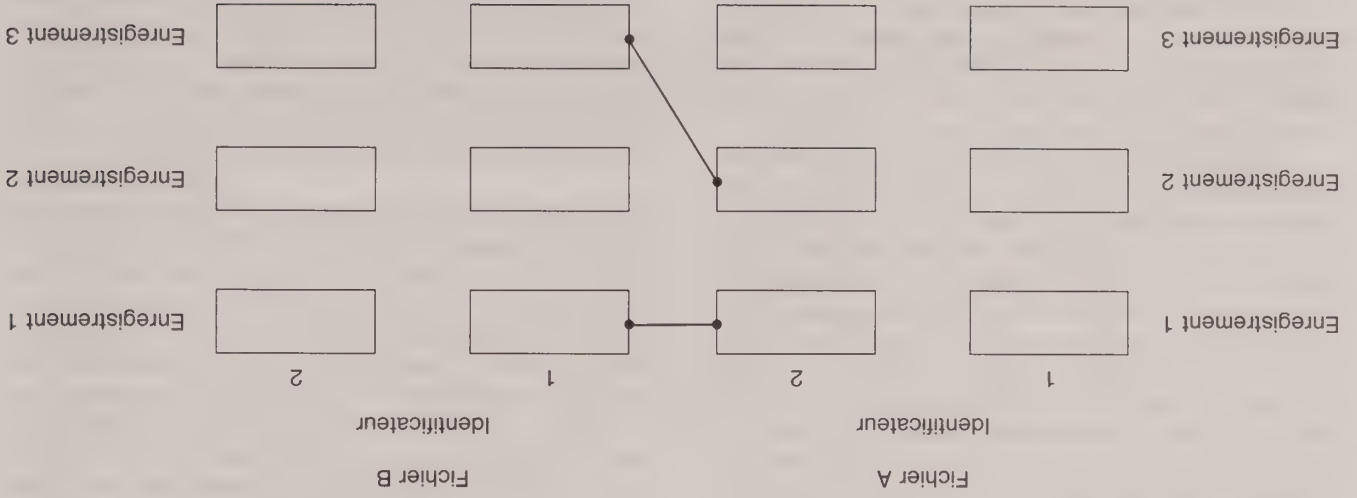


Figure 1. Schéma du couplage de deux fichiers

Evaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisés

S. BARTLETT, D. KREWSKI, Y. WANG et J.M. ZIELINSKI¹

RÉSUMÉ

Les études épidémiologiques qui visent à étudier le rapport entre les risques environnementaux et l'état de santé comptent beaucoup sur l'appariement d'enregistrements de bases de données administratives différentes. Par des algorithmes complexes de couplage d'enregistrements appliqués à de grandes bases de données, on peut évaluer la possibilité d'un appariement de deux enregistrements particuliers en se fondant sur la comparaison d'une ou de plusieurs variables d'identification dans ces enregistrements. Puisque les erreurs d'appariement sont inévitables, il faut pouvoir tenir compte de leur effet sur les inférences statistiques faites à partir des fichiers couplés. Cet article donne un aperçu de la méthodologie utilisée pour le couplage d'enregistrements et traite les questions statistiques qui se rattachent aux erreurs de couplage.

MOTS CLÉS: Couplage d'enregistrements informatisés; étude des exploitants agricoles canadiens; étude de mortalité fondée sur le Fichier dosimétrique national; sélection de seuil.

1. INTRODUCTION

Depuis quelques années, les spécialistes de l'épidémiologie environnementale utilisent de plus en plus des bases de données administratives comme sources de renseignements pour les études sur la santé (Howe et Spasoff 1986; Carpenter et Fair 1990). Grosso modo, cette utilisation consiste à coupler des enregistrements sur l'exposition des personnes aux risques environnementaux avec des enregistrements sur l'état de santé en appariant, souvent par des méthodes informatiques, des enregistrements de bases de données différentes (Newcombe 1988). On a utilisé récemment des méthodes de couplage d'enregistrements informatisés (CEI) pour faire une étude de mortalité sur plus de 326,000 exploitants agricoles du Canada au regard des pratiques agricoles (Jordan-Simpson *et al.* 1990). Dans cette étude, on a couplé la Base canadienne de données sur la mortalité (BCDM) au recensement de la population de 1971 et au recensement de l'agriculture de la même année. Selon des résultats provisoires basés sur un groupe de 70,000 exploitants agricoles de sexe masculin de la Saskatchewan, l'ensemble de la cohorte n'affichait pas de taux de mortalité excessif pour des causes de décès particulières, sauf qu'on notait l'existence d'une relation dose-réponse entre la mortalité due au lymphome non hodgkinien et la superficie traitée aux herbicides pour les exploitations de moins de 1,000 acres (Wigle *et al.* 1990).

Une autre grande étude en cours qui utilise le couplage d'enregistrements est basée sur le Fichier dosimétrique national (FDN) du Canada. Le FDN contient des données sur l'exposition professionnelle au rayonnement ionisant pour environ 255,000 Canadiens; ces données remontent aussi loin que 1950. On a couplé récemment le FDN à la

BCDM afin d'étudier les liens possibles entre l'exposition au rayonnement ionisant et la mortalité par cancer (Ashmore *et al.* 1993).

Il s'est fait d'autres études sur la santé où on a couplé des données sur l'exposition professionnelle à la BCDM. Howe *et al.* (1987) ont observé un nombre significativement élevé de cas de cancer du poumon chez les travailleurs des mines d'uranium dans les Territoires du Nord-Ouest. De même, une étude de cohortes a permis d'établir un rapport significatif entre le cancer du poumon et les fumées des moteurs diesels ou la poussière de charbon chez des retraités de la Compagnie des chemins de fer nationaux du Canada (Howe *et al.* 1983). Shannon *et al.* (1984) ont couplé les fiches d'emploi de travailleurs du nickel de l'Ontario à la BCDM et ont observé chez ce groupe un taux de mortalité excessif pour le cancer du larynx et le cancer du poumon. Morrison *et al.* (1988) ont déterminé un risque significativement élevé de cancer du poumon, des glandes salivaires, de la cavité buccale et du pharynx chez les travailleurs des mines de fluorite de Terre-Neuve. Mao *et al.* (1988) ont eu recours au CEI pour coupler la BCDM au registre du cancer de l'Alberta afin de déterminer les taux de survie après diagnostic pour plusieurs types de cancer. Par ailleurs, on a couplé la base de données de l'Enquête sur la population active du Canada à la BCDM pour faire une étude de mortalité pour diverses professions (Howe et Lindsay 1983). Fair (1989) dresse une liste exhaustive des études sur la santé qui reposent sur le couplage de données sur l'exposition à la BCDM.

Le couplage d'enregistrements est l'opération par laquelle on associe deux ou plusieurs éléments d'information distincts qui se rapportent au même individu. Les méthodes de couplage d'enregistrements informatisés sont devenues très perfectionnées; elles utilisent en effet des

¹ S. Bartlett, D. Krewski, Y. Wang et J.M. Zielinski, Direction de l'hygiène du milieu; Direction générale de la protection de la santé, Santé et Bien-être social Canada, Ottawa (Ontario) Canada K1A 0L2.

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient une section spéciale sur le couplage d'enregistrements et l'appariement statistique. Nous tenons particulièrement à remercier Fritz Scheuren qui a coordonné le travail de rédaction pour cette section spéciale. Un ou deux articles qui portent aussi sur ce sujet et qui ont été reçus trop tard pour être inclus dans le présent numéro pourront être publiés dans un numéro ultérieur.

Dans le couplage d'enregistrements, on combine deux fichiers de données en couplant des enregistrements qui se rapportent à la même unité. L'objectif peut être de créer un fichier de données enrichi renfermant des variables provenant des deux fichiers sources ou de chercher à déterminer des enregistrements se rapportant à des unités communes. Dans les situations où le couplage d'enregistrements n'est pas possible, on pourrait avoir recours à l'appariement statistique pour créer un fichier de données enrichi. Un fichier de données créé par appariement statistique peut renfermer des enregistrements synthétiques en ce sens que les variables obtenues à partir des différentes sources de données n'ont pas à se rapporter à la même unité; cependant, on espère que le fichier apparié reflète toujours de façon exacte les liens statistiques qui existent entre les variables. Bartlett, Krewski, Wang et Zielinski traitent des avantages et des inconvénients du couplage d'enregistrements dans les études épidémiologiques. On étudie les méthodes de couplage d'enregistrements ainsi que des questions méthodologiques qui sont illustrées à l'aide d'exemples tirés de deux études de couplage d'enregistrements à grande échelle en épidémiologie. On examine aussi des questions liées à l'analyse de données provenant de fichiers couplés.

Belin décrit une méthode expérimentale pour évaluer d'autres procédures de couplage d'enregistrements. La méthode est illustrée à l'aide d'une expérience factorielle dans le cadre de laquelle on étudie, entre autres l'effet de facteurs tels que le choix de variables d'appariement et l'affaiblissement des poids. L'expérience fait appel à des données provenant de la répétition générale de 1988 du recensement des États-Unis et de l'enquête postcensitaire correspondante.

Thibaudau examine un modèle autre que le modèle d'indépendance conditionnelle généralement utilisé afin de trouver les probabilités de concordance dans différents champs de comparaison. À titre illustratif, on utilise des données tirées de la répétition générale du recensement et de l'enquête postcensitaire réalisée à St-Louis en 1988. On trouve que le modèle d'indépendance conditionnelle est raisonnable pour les liens véritables; cependant, on utilise un modèle log-linéaire hiérarchique avec termes d'interaction pour les non-liens véritables.

Scheuren et Winkler étudient l'analyse de données provenant de fichiers couplés. En particulier, ils examinent le problème de la régression d'une variable dépendante provenant d'un fichier source par rapport à une variable indépendante provenant d'un autre fichier source. La méthode utilisée consiste à estimer les biais dus à des enregistrements qui auraient pu être couplés incorrectement et à corriger ces biais. Cette méthode donne de bons résultats s'il est possible d'estimer correctement la probabilité qu'un appariement soit un lien véritable (et par conséquent les biais dans l'estimation par régression). Des résultats empiriques sont présentés.

Le dernier article dans cette section spéciale, rédigé par Singh, Mantel, Kinack et Rowe, porte sur l'appariement statistique plutôt que sur le couplage d'enregistrements. Les auteurs élaborent des méthodes d'appariement qui font appel à des données supplémentaires pour éviter de poser l'hypothèse de l'indépendance conditionnelle. Ils considèrent aussi l'imposition de contraintes normales afin que le fichier apparié concorde avec les distributions de variables normales marginales ou conditionnelles appropriées obtenues à partir des fichiers sources ou d'informations supplémentaires. La principale conclusion d'une évaluation empirique est que l'utilisation d'informations supplémentaires appropriées peut améliorer considérablement la qualité du fichier apparié.

Hidiroglou, Drew et Gray présentent des normes pour les définitions de la non-réponse à des enquêtes en voie d'être adoptées à Statistique Canada. Cela facilitera l'analyse de tendances globales dans la non-réponse et une meilleure compréhension des différences dans la non-réponse à différentes enquêtes. On traite aussi des facteurs qui ont une incidence sur la non-réponse et des mesures prises pour réduire cette dernière, et l'on examine les taux de non-réponse pour deux grandes enquêtes de Statistique Canada.

Treder et Sedransk comparent l'échantillonnage aléatoire simple et trois méthodes de répartition pour le double échantillonnage. Ces trois méthodes de répartition sont: la répartition proportionnelle, la répartition de Rao et la répartition optimale.

Casady et Lepkowski proposent des plans d'enquêtes téléphoniques stratifiés basés sur des listes commerciales de numéros de téléphone comme méthodes de rechange pour la méthode de composition aléatoire à deux degrés, largement utilisée et connue sous le nom de technique de Mitofsky-Waksberg. On compare l'efficacité de divers plans d'échantillonnage pour ce plan stratifié, pour la composition aléatoire simple et pour la technique de Mitofsky-Waksberg.

Ouyang, Schreuder, Max et Williams étudient le problème de l'estimation dans l'échantillonnage Poisson-Poisson et binomial-Poisson. On élabore un certain nombre d'estimateurs des totaux et des erreurs-types, et on évalue empiriquement ces estimateurs dans le contexte de l'estimation du volume total de bois utilisable dans un peuplement forestier. Avec le présent numéro, *Techniques d'enquête* passe à un plus grand format de page. Ce format coûte moins cher à imprimer; ce qui permettra à la revue de réduire son déficit permanent de production. Nous avons aussi profité de l'occasion pour remanier la couverture. J'espère que vous aimerez le résultat de nos efforts.

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada
Volume 19, numéro 1, juin 1993

TABLE DES MATIÈRES

Dans ce numéro	1
Couplages d'enregistrements et appariement statistique	
S. BARTLETT, D. KREWSKI, Y. WANG et J.M. ZIELINSKI	
Evaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé	3
T.R. BELIN	
Evaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle	15
Y. THIBAUDEAU	
Le pouvoir discriminant des structures de dépendance dans le couplage d'enregistrements	35
F. SCHEUREN et W.E. WINKLER	
Analyse de régression de fichiers de données couplés par ordinateur	45
A.C. SINGH, H.J. MANTEL, M.D. KINACK et G. ROWE	
Appariement statistique: l'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle	67
M.A. HIDIROGLOU, J.D. DREW et G.B. GRAY	
Cadre pour l'évaluation et la réduction de la non-réponse dans les enquêtes	91
R.P. TREDER et J. SEDRANSK	
Echantillonnage double en vue d'une stratification	107
R.J. CASADY et J.M. LEPKOWSKI	
Plans d'enquête téléphonique stratifiée	115
Z. OUYANG, H.T. SCHREUDER, T. MAX et M. WILLIAMS	
Echantillonnage Poisson-Poisson et binomial-Poisson dans le domaine des forêts	127

TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa
G.J.C. Hole
F. Mayda (Directeur de la production)
M.P. Singh
R. Platek (Ancien président)

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistique Canada*
E.B. Dagum, *Statistique Canada*
J.-C. Deville, *INSEE*
D. Drew, *Statistique Canada*
R.E. Fay, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistique Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*

Rédacteurs adjoints

P. Lavallée, L. Mach et H. Mantel, *Statistique Canada*

POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 \$ par année au Canada, 42 \$ (E.-U.) aux Etats-Unis, et de 49 \$ (E.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.



Ottawa

ISSN 0714-0045

N° 12-001 au catalogue

Autres pays : 49 \$ US

États-Unis : 42 \$ US

Prix : Canada : 35 \$

Juillet 1993

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, par quelque moyen que ce soit, électronique, mécanique, photographique, magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division de la commercialisation, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, des Sciences
et de la Technologie, 1993

Publication autorisée par le ministre
responsable de Statistique Canada

JUIN 1993 • VOLUME 19 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

Ans Years of
d'excellence Excellence



TECHNIQUES D'ENQUÊTE



NUMÉRO 1

VOLUME 19

JUIN 1993

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

Catalogue 12-001

TECHNIQUES D'ENQUÊTE



12-001

